Discrete Language Models for Video Retrieval

Kieran Mc Donald

Bachelor of Science in Computer Applications

A dissertation submitted in fulfillment
of the requirements for the award of
Doctor of Philosophy (Ph.D.)

Supervisor: Prof. Alan F. Smeaton

School of Computing
Dublin City University

September 2005

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work

Signed Kieran McDonald
ID No 95668292
Date 29/9/2005

To my father Richie and

in memory of my mother Mary

ACKNOWLEDGEMENTS

Firstly, I must thank my supervisor Alan Smeaton for his guidance and support I don't know where he gets all the energy from but I am glad he was there for me when I needed his help I am indebted to the support of the School of Computing and Enterprise Ireland for funding my research under different projects over the years

I wish to thank my family, who are as relieved as I am that I have finally finished and without their support in life I would never have got this far

I need to thank Paul Browne and Jiamin Ye (Mr & Mrs Browne) for their friendship and company at coffee through the years I especially thank Wu Hai for being a great friend and for the unforgettable adventures around Europe and China

I am grateful to my ex-college mates from undergraduate years, Mike, Aidan, Brian, Jen, Mel and also Steve who remain good friends The poker sessions over the last couple of years have provided eventful nights and sometimes early mornings

Lastly, I would like to thank all those people from CDVP and DCU who made college life fun - Cathal, Tom, Hyowon, Jer, Qamir, Orla, Georgina, Pete, Neil, Aoife, Mary, Michelle, Nano, Paul Ferguson, Fabrice, Colum, Sandra, Sinead, James, Mike, Dalen, Tibo, Nico, Chris and of course my ex-flatmates Irene, Lubos, Yan and Kasie

TABLE OF CONTENTS

DEDICATION iii

ACKNOWLEDGEMENTS iv

LIST OF TABLES xi

LIST OF FIGURES xviii

SUMMARY xxiii

I INTRODUCTION 1

1 1 Video Retrieval 1

1 2 Discrete Language Modelling approach to Video Retrieval 3

1 3 Research Objectives 4

1 4 Thesis Organisation 6

II INFORMATION RETRIEVAL USING LANGUAGE MODELS 9

2 1 Introduction 9

2 2 A Simple Language Model for IR 10

2 3 Probability Models 13

2 4 Statistical Estimation 14

2 4 1 Discounting-Based Smoothing 15

2 4 2 Combination-Based Smoothing 17

2 4 3 Back-Off Smoothing 19

2 4 4 Validation-based Smoothing 19

2 5 Ranking Documents 19

2 5 1 Query-Likelihood 21

2 5 2 Document-Likelihood 22

2 5 3 Relative Entropy and Cross Entropy 22

2 6 Relevance Feedback 23

2 6 1 Query-Likelihood Feedback 24

2 6 2 Query Model Feedback 25

2 7 Non-Language modelling approaches to Information Retrieval 27

2 7 1 Boolean Model 27

2 7 2 Vector Space Model 28

2 7 3 Probabilistic Models 30

2 8 Summary 35

III STATE-OF-THE-ART IN VIDEO RETRIEVAL 38

3 1 Introduction 38

3 2 Video Indexing 40

3 2 1 Structuring Video 44

3 2 2 Text Descriptions 46

3 2 3 High-Level Concepts 47

3 2 4 Visual Descriptions 49

3 2 5 Audio Descriptions 68

3 3 Video Retrieval Models 69

3 3 1 Query Preprocessing 70

3 3 2 Text-Based Video Retrieval 71

3 3 3 High-Level Concept-Based Video Retrieval 72

3 3 4 Visual-Based Video Retrieval 73

3 3 5 Fusion 80

3 3 6 Relevance Feedback 83

3 4 Evaluation of Video Retrieval using TRECVid 85

3 4 1 Reference Collection 86

3 4 2 Search Topics 87

3 4 3 Relevance Judgements 91

3 4 4 Search Experiments 92

3 4 5 Evaluation Measures 93

3 5 Summary 96

IV PROPOSED DISCRETE LANGUAGE MODELLING APPROACH FOR VIDEO RETRIEVAL 99

4 1 Introduction 99

4 2 Proposed Extensions to the Text-based Hierarchical Language Model for Video Shot Retrieval 101

4 3 Proposed Visual Language Models 103

4 3 1 Visual Colour Language 105

4 3 2 Visual Edge Language 106

4 3 3 Visual Texture Language 107

4 4 Fusion Methods 108

4 4 1 Combining Multiple Features 110

4 4 2 Combining Multiple Visual Examples 111

4 4 3 Combining Text and Visual 112

4 5 Evaluation Methodology 113

4 6 Related Research 115

4 7 Summary 117

V EVALUATION I VIDEO RETRIEVAL USING TEXT FEATURES 118

5 1 Introduction 118

5 2 Experiments with non-hierarchical structures 120

5 3 Experiments with hierarchical physical video structures 127

5 4 Experiments with hierarchical semantic video structure 134

5 5 Summary 139

VI EVALUATION II VIDEO RETRIEVAL USING VISUAL FEATURES 141

6 1 Introduction 141

6 2 Overview of Experiments 142

6 3 Experiments with Colour Features 143

6 3 1 Global Colour 143

6 3 2 Regional Colour 146

6 3 3 HSV colour experiments using the physical and semantic video structure 150

6 4 Experiments with Edge Features 158

6 4 1 Global Canny Edge 160

6 4 2 Regional Canny Edge 163

6 5 Experiments with Texture Feature 168

6 5 1 Global Texture 170

6 5 2 Regional Texture 178

6 6 Summary 186

VII EVALUATION III VIDEO RETRIEVAL USING COMBINED FEATURES 188

7 1 Introduction 188

7 2 Experiments with multiple visual features 189

7 3 Experiments with multiple visual examples 195

7 4 Experiments with multiple modalities 203

7 5 Discussion 215

7 6 Summary 217

VIII CONCLUSIONS 218

8 1 General Conclusions 218

8 2 Text Language Models 221

8 3 Visual Language Models 223

8 4 Fusion Methods 225

8 5 Evaluation Methodology 227

8 6 Future Work 228

8 7 Summary 230

APPENDIX A — TRECVID TOPICS 231

APPENDIX B — ADDITIONAL TABLES FOR CHAPTER 5 235

APPENDIX C — ADDITIONAL TABLES FOR CHAPTER 6 247

APPENDIX D — ADDITIONAL TABLES FOR CHAPTER 7 264

APPENDIX E — PAPERS PUBLISHED ON THIS WORK 280

REFERENCES 281

LIST OF TABLES

1 TRECVid High-level features 48

2 Visual and multimodal results for the continuous GMM DCT query-likelihood approach for the TRECVid search tasks 79

3 Visual and multimodal results for the continuous GMM DCT query-likelihood, the discrete multinomial DCT query-likelihood and the continuous DCT document-likelihood approaches for the TRECVid 2003 search task 79

4 Distribution of topics within each topic classification type for the TRECVid search tasks 89

5 Summary statistics of the visual examples and relevant shots for the TRECVid search tasks 91

6 Additional query stopwords for the TRECVid search topics 119

7 Comparison of ASR *shot-only*, *adj-only*, *video-only* and *story-only* structures on TRECVid 2002-2004 121

8 Statistical significance comparison of retrieval models on the ASR *adj-only* structure for the aggregated TRECVid 2002-2004 search task 125

9 Comparison of retrieval models on the ASR *shot+adj+vid* structure with the *shot-only*, *adj-only*, *story-only* and *shot+video* structures for TRECVid 2002–2004 132

10 Comparison of retrieval models on the ASR *shot+story* and *shot+adj+story* structures with the *shot-only* and *story-only* structures for TRECVid 2003 136

11 Statistical significance comparison of retrieval models on the regional 5x5 HSV 16x4x4 colour feature for the aggregated TRECVid 2002, 2003 and 2004 search tasks 150

12 Comparison of retrieval models on the HSV 5x5 16x4x4 feature for official TRECVid topics (i e fused topic examples) 151

13 Results for smoothing the HSV 80x1x1+1 *shot+video* hierarchical colour structure for the *TRECVid 2002* search task 158

14 Results for smoothing the HSV 80x1x1+1 *shot+adj+video* hierarchical colour structure for the *TRECVid 2002* search task 159

15 Statistical significance comparison of retrieval models on the regional 5x5 Canny 64+1 edge feature for the aggregated TRECVid 2002, 2003 and 2004 search tasks 168

16 Comparison of retrieval models on the Canny 5x5 64+1 feature for official TRECVid topics (i e fused topic examples) 169

17 Statistical significance comparison of retrieval models on the regional 5x5 DCT 3x3x3x3x3 texture feature for the aggregated TRECVid 2002, 2003 and 2004 search tasks 184

18 Comparison of retrieval models on the DCT 5x5 3x3x3x3x3 feature for official TRECVid topics (i e fused topic examples) 185

19 Comparison of the average results across retrieval models for combining the colour, edge and texture results using the *Vis-CombJointPr*, *Vis-CombWtRank* and *Vis-CombWtScore* fusion methods on the TRECVid search tasks 190

20 Comparison of retrieval models for the TRECVid 2002-2004 search tasks using the *Vis-CombWtScore* fusion method for combining visual features 194

21 Comparison of the average results across retrieval models for combining visual examples using different fusion methods for TRECVid 2002-2004 196

22 Comparison of unbiased retrieval models for the TRECVid 2002, 2003 and 2004 search tasks using the *VisExs-CombMaxScore* fusion method for combining colour, edge and texture results 201

23 Comparison of the average results across retrieval models for combining the *shot+adj+video* interpolated text language model's results with the combined visual examples (*VisExs-CombMaxScore*) results of the different retrieval models 204

24 Comparison of the text-only, visual-only and fusion methods *TextVis-CombJointPr*, *TextVis-CombWtRank* and *TextVis-CombWtScore* for combining the text results with the Jelinek-Mercer language model's visual results on the TRECVid 2002, TRECVid 2003 and TRECVid 2004 collections 209

25 Comparison of unbiased retrieval models for the TRECVid 2002, 2003 and 2004 search tasks using the *TextVis-CombMaxScore* fusion method for combining text and results 210

26 Comparison of the average results across retrieval models for combining the *Shot+Adj+Video* interpolated text language model's results with the combined visual examples (*VisExs-CombMaxScore*) results of the different retrieval models using the *TextVisBoth-Oracle* and *TextVisComb-Oracle* for the TRECVid search tasks 213

27 Comparison of unbiased retrieval models for the TRECVid 2002, 2003 and 2004 search tasks for the *Vis-CombWtScore*, *VisExs-CombMaxScore* and *TextVis-CombMaxScore* fusion methods 215

28 TRECVid 2002 search topics 232

29 TRECVid 2003 search topics 233

30 TRECVid 2004 search topics 234

31 ASR *adj-only* versus *shot-only* and *video-only* for TRECVid 2002 236

32 ASR *adj-only* versus *shot-only*, *video-only* and *story-only* for TRECVid 2003 237

33 ASR *adj-only* versus *shot-only* and *video-only* for TRECVid 2004 238

34 ASR *story-only* versus *shot-only*, *adj-only* and *video-only* for TRECVid 2003 239

35 Statistical significance comparison of retrieval models for the ASR *adj* representation on TRECVid 2002 240

36 Statistical significance comparison of retrieval models for the ASR *adj* representation on TRECVid 2003 241

37 Statistical significance comparison of retrieval models for the ASR *adj* representation on TRECVid 2004 242

38 Statistical significance comparison of retrieval models for the ASR *story* representation on TRECVid 2003 243

39 Statistical significance comparison of hierarchical LMs for the ASR *shot+video* representation on TRECVid 2002–2004 244

40 Statistical significance comparison of hierarchical LMs for the ASR *shot+adj+video* representation on TRECVid 2002–2004 245

41 Statistical significance comparison of hierarchical LMs for the ASR *shot+story* representation on TRECVid 2003 246

42 Statistical significance comparison of hierarchical LMs for the ASR *shot+adj+story* representation on TRECVid 2003 246

43 Comparison of retrieval models on the *global HSV 16x4x4* feature for the *TRECVid 2002* and *TRECVid 2003* search tasks 248

44 Comparison of retrieval models on the regional *HSV 5x5 16x4x4* feature for the *TRECVid 2002* and *TRECVid 2003* search task 249

45 Comparison of the *indexing units* shots, sequence of adjacent shots, and videos with the HSV 80x1x1+1 colour representation for language models and visual retrieval models on the *TRECVid 2002* search task 250

46 Results for using the *indexing unit* sequence of adjacent shots with the HSV 80x1x1+1 colour representation for language models and visual retrieval models on the *TRECVid 2003* search task 251

47 Comparison of *global Canny 64+1 edge representations* with *Canny 4+1*, *Canny 16+1* and *Canny 32+1* representations for language models and standard visual retrieval models for the *TRECVid 2002* search task 252

48 Comparison of *global Canny 64+1 edge representations* with *Canny 4+1*, *Canny 16+1* and *Canny 32+1* representations for language models and standard visual retrieval models for the *TRECVid 2003* search task 253

49 Comparison of *regional Canny edge representations* (Canny 64+1 for 3x3 regions) using language models and standard visual retrieval models for the *TRECVid 2002* search task 254

50 Comparison of *regional Canny edge histogram representations* (Canny 64+1 for 5x5 regions) using language models and standard visual retrieval models for the *TRECVid 2003* search task 255

51 Comparison of *global DCT 4x4x4x4 representation* with the DCT 8x8x8 and DCT 3x3x3x3x3 representations using language models and standard visual retrieval models for the *TRECVid 2002* search task 256

52 Comparison of *global DCT 8x8x8 representation* with the DCT 3x3x3x3x3 representation using language models and standard visual retrieval models for the *TRECVid 2003* search task 257

53 Comparison of language models and standard visual retrieval models using the *global DCT 3x3x3x3x3 representation* for the *TRECVid 2002* and *TRECVid 2003* search tasks 258

54 Comparison of the *5x5 regional DCT 3x3x3x3x3 representations* with the non-regional texture representation using language models and standard visual retrieval models for the *TRECVid 2002* search task 259

55 Comparison of the *5x5 regional DCT 3x3x3x3x3 texture representations* with the non-regional and regional 3x3 and 4x4 representations using language models and standard visual retrieval models for the *TRECVid 2003* search task 260

56 Comparison of unbiased retrieval models for the TRECVid 2002 search tasks 5x5 regional HSV colour, Canny edge and DCT texture 261

57 Comparison of unbiased retrieval models for the TRECVid 2003 search tasks 5x5 regional HSV colour, Canny edge and DCT texture 262

58 Comparison of unbiased retrieval models for the TRECVid 2004 search tasks 5x5 regional HSV colour, Canny edge and DCT texture 263

59 Comparison of the *Vis-CombWtScore* fusion of the colour, edge and texture results with the colour-only results and the other fusion methods *Vis-CombJointPr* and *Vis-CombWtRank* on the TRECVid 2002 collection 265

60 Comparison of the *Vis-CombWtScore* fusion of the colour, edge and texture results with the colour-only results and the other fusion methods *Vis-CombJointPr* and *Vis-CombWtRank* on the TRECVid 2003 collection 266

61 Comparison of the *Vis-CombWtScore* fusion of the colour, edge and texture results with the colour-only results and the other fusion methods *Vis-CombJointPr* and *Vis-CombWtRank* on the TRECVid 2004 collection 267

62 Comparison of unbiased retrieval models for the TRECVid 2002 search tasks using the *Vis-CombWtScore* fusion method for combining colour, edge and texture results 268

63 Comparison of unbiased retrieval models for the TRECVid 2003 search tasks using the *Vis-CombWtScore* fusion method for combining colour, edge and texture results 268

64 Comparison of unbiased retrieval models for the TRECVid 2004 search tasks using the *Vis-CombWtScore* fusion method for combining colour, edge and texture results 269

65 Comparison of the *VisExs-CombMaxScore* fusion method with the other fusion methods *VisExs-CombJointPr*, *VisExs-CombRank*, *VisExs-CombScore* and *VisExs-CombMaxRank* for combining the visual examples' results of the different retrieval models on the TRECVid 2002 collection 270

66 Comparison of the *VisExs-CombMaxScore* fusion method with the other fusion methods *VisExs-CombJointPr*, *VisExs-CombRank*, *VisExs-CombScore* and *VisExs-CombMaxRank* for combining the visual examples results of the different retrieval models on the TRECVid 2003 collection 271

67 Comparison of the *VisExs-CombMaxScore* fusion method with the other fusion methods *VisExs-CombJointPr*, *VisExs-CombRank*, *VisExs-CombScore* and *VisExs-CombMaxRank* for combining the visual examples' results of the different retrieval models on the TRECVid 2004 collection 272

68 Comparison of unbiased retrieval models for the TRECVid 2002 search tasks using the *VisExs-CombMaxScore* fusion method for combining visual examples 273

69 Comparison of unbiased retrieval models for the TRECVid 2003 search tasks using the *VisExs-CombMaxScore* fusion method for combining visual examples 273

70 Comparison of unbiased retrieval models for the TRECVid 2004 search tasks using the *VisExs-CombMaxScore* fusion method for combining visual examples 274

71 Comparison of the *TextVis-CombWtScore* text and visual fusion method with the text-only results and the other fusion methods *TextVis-CombJointPr* and *TextVis-CombWtRank* for combining the *Shot+Adj+Video* interpolated text language model's results with the combined visual examples results of the different retrieval models for the TRECVid 2002 collection 275

72 Comparison of the *TextVis-CombWtScore* text and visual fusion method with the text-only results and the other fusion methods *TextVis-CombJointPr* and *TextVis-CombWtRank* for combining the *Shot+Adj+Video* interpolated text language model's results with the combined visual examples results of the different retrieval models for the TRECVid 2003 collection 276

73 Comparison of the *TextVis-CombWtScore* text and visual fusion method with the text-only results and the other fusion methods *TextVis-CombJointPr* and *TextVis-CombWtRank* for combining the *Shot+Adj+Video* interpolated text language model's results with the combined visual examples results of the different retrieval models for the TRECVid 2004 collection 277

74 Comparison of unbiased retrieval models for the TRECVid 2002 search tasks using the *TextVis-CombWtScore* fusion method for combining text and visual results 278

75 Comparison of unbiased retrieval models for the TRECVid 2003 search tasks using the *TextVis-CombWtScore* fusion method for combining text and visual results 278

76 Comparison of unbiased retrieval models for the TRECVid 2004 search tasks using the *TextVis-CombWtScore* fusion method for combining text and visual results 279

LIST OF FIGURES

1	HMM approach to IR	14
2	LMIR using query-likelihood	20
3	LMIR use relative entropy	20
4	Overview of video indexing system	41
5	Overview of the MPEG7 Multimedia description schemes	43
6	Segmentation of the audiovisual content	44
7	Illustration of the different segmentations of a television news programme	45
8	Examples of different shot transitions	45
9	Spatial localisation of visual features	49
10	Brodatz Texture Examples	55
11	Gabor filters	57
12	Frequency layout for the MPEG7 Homogeneous Texture descriptor	58
13	DCT basis functions	59
14	MPEG7 Edges	61
15	Camera motion	63
16	Overview of video retrieval system	69
17	TRECVid 2002 search topic example	88
18	TRECVid 2003 search topic example	89
19	Precision and Recall	93
20	Global and regional whole image representations	105

21 Canny edge feature 107

22 Fusion of retrieval models for TRECVid topics 109

23 Comparison of retrieval models on the non-hierarchical ASR structures for TRECVid 2002-2004 122

24 Graph of MAP versus parameter values for LMs on the ASR *adj-only* and *story-only* structure for TRECVid 2002-2004 128

25 Comparison of topic results for Dirichlet LM on the non-hierarchical ASR structures for TRECVid 2002-2004 129

26 Comparison of LMs on the ASR *shot+video* and *shot+adj+video* hierarchical physical structures with the *adj-only* structure for TRECVid 2002-2004 130

27 Comparison of Jelinek-Mercer LM topic results on the ASR *shot+video*, *shot+adj+video*, *shot-only* and *adj-only* structures for TRECVid 2002-2004 133

28 Comparison of hierarchical LMs on the ASR *shot+story*, *shot+adj+story* hierarchical semantic structures with the *story-only* and *shot+adj+video* structures for TRECVid 2003 135

29 Comparison of Jelinek-Mercer LM topic results for the ASR *story-only*, *shot+story*, *shot+adj+story* and *shot+adj+video* structures on TRECVid 2003 138

30 Comparison of *global HSV histogram representations* (H 80 +1 desaturated, HSV 5x5x5, HSV 16x4x4) using language models and standard visual retrieval models for the *TRECVid 2002* and *TRECVid 2003* search task 144

31 Comparison of *regional histogram representations* (HSV 16x4x4 for no regions, 3x3, 4x4, and 5x5 regions) using language models and standard visual retrieval models for the *TRECVid 2002* and *TRECVid 2002* search tasks 147

32 Plot of MAP over the parameter space for the parametric language models using regional colour representations for the *TRECVid 2002* search task and *TRECVid 2003* search task 149

33 Comparison of the *indexing units* shots, sequence of adjacent shots, and videos with the HSV 80x1x1+1 colour representation for language models and visual retrieval models on the *TRECVid 2002* search task 152

34 Comparison of the *indexing units* shots, sequence of adjacent shots, videos, and stories with the HSV 80x1x1+1 colour representation for language models and visual retrieval models on the *TRECVid 2003* search task 153

35 Comparison of the *indexing units* shots, sequence of adjacent shots, videos and stories with the HSV 80x1x1+1 colour representation on the 20 most successful *TRECVid 2003* topic images using the best Lidstone language models 154

36 Comparison of the *indexing units* shots, sequence of adjacent shots, videos and stories with the HSV 80x1x1+1 colour representation on the 20 most successful *TRECVid 2003* topic images using the best Lidstone language models 154

37 Comparison of *smoothing with structural units* (shots+videos, shots+sequence of adjacent shots+videos) and *smoothing different indexing units* (shot-only, video-only, adjacent shots-only) with the HSV 80x1x1+1 colour representation for interpolation-based language modelling information retrieval on the *TRECVid 2002* search task 156

38 Comparison of *smoothing with structural units* (shots+videos, shots+adjacent shots+videos) and *smoothing different indexing units* (shot-only, video-only, adjacent shots-only) with the HSV 80x1x1+1 colour representation for interpolation-based language modelling information retrieval on the *TRECVid 2003* search task 156

39 Comparison of *smoothing with structural units* (shots+videos, shots+adjacent shots+videos) and *smoothing different indexing units* (shot-only, video-only, adjacent shots-only) with the HSV 80x1x1+1 colour representation on the 20 most successful *TRECVid 2002* topic images using the best Jelinek-Mercer structural smoothing model 157

40 Comparison of *global Canny edge representations* with different number of bins (4, 16, 32, 64) using language models and standard visual retrieval models for the *TRECVid 2002* search task 161

41 Comparison of *global Canny edge representations* with different number of bins (4, 16, 32, 64) *on the 20 most successful TRECVid 2002 topic images* using the optimised Absolute discounting language model 162

42 Comparison of *global Canny edge representations* with different number of bins (4, 16, 32, 64) *on the 20 most successful TRECVid 2003 topic images* using the optimised Lidstone language model 162

43 Comparison of *regional Canny edge representations* (Canny 64+1 for no regions, 3x3, 4x4, and 5x5 regions) using language models and standard visual retrieval models for the *TRECVid 2002* and *TRECVid 2003* search tasks 164

44 Comparison of *regional Canny edge representations* (Canny 64+1 for no regions, 3x3, 4x4, and 5x5 regions) *on the 20 most successful TRECVid 2002 topic images* using the optimised Absolute interpolation language model 165

45 Comparison of *regional Canny edge representations* (Canny 64+1 for no regions, 3x3, 4x4, and 5x5 regions) *on the 20 most successful TRECVid 2003 topic images* using the optimised Linear discounting language model 165

46 Plot of MAP over the parameter space for the parametric language models using regional Canny edges for *TRECVid 2002* search task and *TRECVid 2003* search task 167

47 Comparison of *global DCT representations* (DCT 10x10, DCT 8x8x8, DCT 4x4x4x4, and DCT 3x3x3x3x3) using language models and standard visual retrieval models for the *TRECVid 2002* search task over (a) all topics and (b) all but topic 77 171

48 Comparison of *global DCT representations* (DCT 10x10, DCT 8x8x8, DCT 4x4x4x4 and DCT 3x3x3x3x3) *on the 20 most successful TRECVid 2002 topic images* using the optimised Absolute interpolation language model 172

49 Comparison of MAP for the different parameter values of the parametric language models (Lidstone, Linear, Absolute discounting, Jelinek-Mercer and Absolute interpolation) using the global DCT texture representations (DCT 10x10, DCT 8x8x8, DCT 4x4x4x4 and DCT 3x3x3x3x3) for the *TRECVid 2002* search task for all topics and for all but topic 77 173

50 Comparison of *global DCT representations* (DCT 10x10, DCT 8x8x8, DCT 4x4x4x4 and DCT 3x3x3x3x3) using language models and standard visual retrieval models for the *TRECVid 2003* search task 175

51 Comparison of *global DCT representations* (DCT 10x10, DCT 8x8x8, DCT 4x4x4x4 and DCT 3x3x3x3x3) *on the 20 most successful TRECVid 2003 topic images* using the optimised Jelinek-Mercer language model 175

52 Graph of MAP Response of parametric LMs for global DCT feature 176

53 Comparison of *regional DCT representations* (DCT 3x3x3x3x3 for no regions, 3x3, 4x4 and 5x5 regions) using language models and standard visual retrieval models for the *TRECVid 2002* search task over (a) all topics and (b) all but topic 77 179

54 Comparison of *regional DCT representations* (DCT 3x3x3x3x3 for no regions, 3x3, 4x4 and 5x5 regions) *on the 20 most successful TRECVid 2002 topic images* using the optimised Absolute interpolation language model 180

55 Comparison of *regional DCT representations* (DCT 3x3x3x3x3 for no regions, 3x3, 4x4 and 5x5 regions) using language models and standard visual retrieval models for the *TRECVid 2003* search task 181

56 Comparison of *regional DCT representations* (DCT 3x3x3x3x3 for no regions, 3x3, 4x4 and 5x5 regions) *on the 20 most successful TRECVid 2003 topic images* using the optimised Jelmek-Mercer language model 181

57 Comparison of MAP for the different parameter values of the parametric language models (Lidstone, Linear, Absolute discounting, Jelinek-Mercer and Absolute interpolation) using the *regional DCT representations* (DCT 3x3x3x3x3 for no regions, 3x3, 4x4 and 5x5 regions) for the *TRECVid 2002* and *TRECVid 2003* search tasks 183

58 Comparison of visual feature fusion methods for TRECVid 2002–2004 192

59 Comparison of fusion methods for visual examples on TRECVid 2002–2004 197

60 Comparison of topic results of fusion methods for visual examples of Jelinek-Mercer LM on TRECVid 2002–2004 200

61 Comparison of *VisExs-CombMaxScore* fusion of visual examples with the single best visual example per topic for TRECVid 2002–2004 201

62 Comparison of fusion methods for combining text and visual results on TRECVid 2002–2004 205

63 Comparison of fusion methods for text and visual results of the Jelmek-Mercer LM on individual topics in TRECVid 2002–2004 208

64 Comparison of Oracle results for combining text and visual results on TRECVid 2002–2004 214

Discrete Language Models for Video Retrieval

Kieran Mc Donald

Abstract

Finding relevant video content is important for producers of television news, documentaries and commercials As digital video collections become more widely available, content-based video retrieval tools will likely grow in importance for an even wider group of users In this thesis we investigate language modelling approaches, that have been the focus of recent attention within the text information retrieval community, for the video search task Language models are smoothed discrete generative probability distributions generally of text and provide a neat information retrieval formalism that we believe is equally applicable to traditional visual features as to text We propose to model colour, edge and texture histogram-based features directly with discrete language models and this approach is compatible with further traditional visual feature representations We provide a comprehensive and robust empirical study of smoothing methods, hierarchical semantic and physical structures, and fusion methods for this language modelling approach to video retrieval The advantage of our approach is that it provides a consistent, effective and relatively efficient model for video retrieval

CHAPTER I

INTRODUCTION

Finding relevant video content is important for producers of television news, documentaries and commercials As digital video collections become more widely available, content-based video retrieval tools will likely grow in importance for an even wider group of users In this thesis we investigate language modelling approaches, that have been the focus of recent attention within the text information retrieval community, for the video search task Language models are smoothed discrete generative probability distributions generally of text and provide a neat information retrieval formalism that we believe is equally applicable to traditional visual features as to text We propose to model colour, edge and texture histogram-based features directly with discrete language models and this approach is compatible with further traditional visual feature representations We provide a comprehensive and robust empirical study of smoothing methods, hierarchical semantic and physical structures, and fusion methods for this language modelling approach to video retrieval The advantage of our approach is that it provides a consistent, effective and relatively efficient model for video retrieval

1 1 Video Retrieval

1 2 Discrete Language Modelling approach to Video Retrieval

1 3 Research Objectives

1 4 Thesis Organisation

1.1 Video Retrieval

Video search tools are important in professional video archives since it is far more cost-efficient to re-use content than to reshoot, indeed when dealing with content of a historical nature it may be impossible to reshoot Producers of news, documentaries and commercials therefore have a direct need for effective video content search tools The wider availability of the internet has lead to a change in how people access and find predominantly text-based information and a similar change in peoples' interactions with video may occur when online access to video collections becomes more widespread within society

Current operational video retrieval systems depend heavily on manual indexing in order to provide effective content organisation and search indexes to enable their users to locate video content The manual indexing process is costly, time-consuming and depends heavily on consistently assigning keywords to video content In practice keyword ontologies change over time and inconsistencies develop making retrieval by end-users even more difficult than initially learning the indexing language Exhaustive manual indexing of a video's topical and visual content is not possible and therefore archivists selectively index the most important characteristics of the content Problems occur with this approach as what is important for the end-user may not have seemed significant at indexing time Further difficulties arise when the searcher is interested in visual characteristics since these are often hard to express and index in terms of keywords A key advantage of manual indexing is that descriptions of contextual information and visual content is at a higher semantic level than is possible with current automatic content-based video retrieval systems

Content-based video retrieval systems provide an alternative to video search systems that are based on manual indexes Content-based video retrieval systems automatically index video material by segmenting it into clips and extracting features such as text, colour, texture, motion from each clip to support search These systems provide access to the content via full text search, visual query-by-example and in some cases sketching tools that may support the specification of colour, texture and motion attributes of sought after visual content These content-based search tools are often integrated with video browsing and playback to support efficient navigation and previewing of content within the video retrieval system Manual and fully automatic video indexing systems need not be separated and can be integrated to produce a semi-automatic indexing system that combines the best of both techniques to video retrieval users

In the early 90s research on content-based image retrieval began in earnest to facilitate the search of image collections by their content without the need for costly manual annotation (Flickner et al , 1995, Bach et al , 1996, Pentland et al , 1996, Smith and Chang, 1996b, Rui, Huang and Chang, 1997) Soon afterwards research began on content-based video retrieval systems with some of the early image retrieval systems being adapted and other new video retrieval systems being developed (Ortega et al , 1997, Chang et al , 1997, Wactlar, 2000) The early video retrieval research lacked the rigour of controlled retrieval experiments on common test collections and search tasks that is the backbone of much research within the text information retrieval community Early research on the visual retrieval methods was often performed on collections of images that fit into homogenous categories and is therefore unrepresentative of the video retrieval problem

In the last couple of years there has been a concerted effort in the video retrieval research community through the TRECVid initiative to perform controlled video retrieval

experiments on reasonable sized common video collections and search tasks (Smeaton et al , 2002, Smeaton and Over, 2003, Smeaton et al , 2004b, Kraaij et al , 2004, Smeaton et al , 2004a) The TRECVid search topics consist of general and specific visually oriented requests for people, things, locations and events and seek to replicate the type of requests that are common in professional video archives Each search topic is a multimedia description of the video need containing a text description and multiple image and video examples TRECVid provides a firmer basis on which to investigate content-based video retrieval than was previously possible in controlled and repeatable experimental conditions

1.2 *Discrete Language Modelling approach to Video Retrieval*

Language models are discrete generative probability distributions of text that were originally developed in the speech recognition community and have been the focus of recent attention within the text information retrieval community The standard language modelling approach to information retrieval models each document using a discrete generative probability distribution and ranks documents by their language model's probability of generating the query text (Ponte and Croft, 1998) If documents are represented by their empirical probability distribution (i e based on relative frequency of terms) a problem occurs for all documents missing one or more of the query terms as they will have the same probability of zero of generating the query, which is undesirable within an information retrieval system as this produces a poor ranking of documents Many different smoothing techniques have been developed for speech recognition and later for text information retrieval to handle this so-called zero frequency problem by adjusting or smoothing in some way the empirical distribution

We believe that the language modelling approach to information retrieval is equally applicable to traditional visual features as to text We propose to model visual features, namely colour, edge and texture histograms, directly with language models (discrete generative probability distributions) This approach is not only compatible with visual histogram representations but also with further standard visual feature representations such as co-occurrence matrices and correlograms It can also be applied to some of the standard MPEG7 Multimedia Content Description Interface features Because low-level visual features are semantically so different to text, we believe it is necessary to investigate the different smoothing techniques in this new context

Our approach differs from current video retrieval approaches, which typically use geometric distances such as Euclidean distance or Manhattan distance to perform matching on these types of visual features (Hauptmann et al , 2004, Pickering et al , 2003, Rautiainen, Penttila, Vorobiev, Noponen, Vayrynen, Hosio, Matinmikko, Makela, Peltola,

Ojala and Seppanen, 2003, Snoek et al , 2005) It differs from the current generative probabilistic approach to video retrieval (Westerveld, de Vries and van Ballegooij, 2003), which models a single visual feature, multi-spectral DCT coefficients, using a continuous probability Gaussian Mixtures Model (GMM) that is estimated with the iterative EM algorithm (Dempster et al , 1977) and smoothed by the feature's marginalised distribution over the collection Our discrete language modelling approach for video retrieval is quicker to index and retrieve than this continuous language model approach but cannot handle high-dimensional features as efficiently We refer to our model as a *discrete language model* aware of the inherent redundancy in such a phrase but to distinguish it from the GMM language modelling approach to video retrieval

The advantage of our approach is that it provides a consistent, effective and relatively efficient model for video retrieval This approach is consistent as we transform the visual features colour, edge and texture into a language of terms and model them in much the same way by using language models as for our text feature This approach is efficient as both in text and visual retrieval the time complexity of the language modelling approach is not greater than traditional retrieval models in both mediums We also show in this thesis that this approach is as effective as others for the video retrieval task

We evaluate our approach on three video retrieval test collections, TRECVid 2002, TRECVid 2003 and the recent TRECVid 2004 collections Through these experiments we provide a comprehensive and robust empirical study of smoothing methods, hierarchical semantic and physical structures, and fusion methods for this discrete language modelling approach to video retrieval

1.3 Research Objectives

The main objective of our research is to thoroughly investigate language modelling approaches for text and visual-based video retrieval Due to the multimodal nature of video retrieval, we also take on a subsidiary research objective of investigating fusion models for combining the language modelling retrieval results of different features

We provide a robust empirical evaluation of these retrieval methods by testing on multiple video retrieval test collections to increase the number of topics, by using unbiased, controlled and repeatable experimental setups, by comparing against multiple standard retrieval models to provide a credible non-language modelling baseline and by using statistical significance tests to discern credible differences in performance from differences due to chance It is only recently with the availability of the TRECVid video search test collections that such robust empirical studies of video retrieval can be attempted

For text-based video shot retrieval we are interested in answering the following research questions

- How effective is the language modelling approach compared to traditional retrieval models such as TF-IDF and BM25?
- Do semantic structures such as story units dramatically improve the performance of video shot retrieval?
- Do alternative smoothing methods improve on the text language models that use Jelinek-Mercer interpolation smoothing?

For visual-based video shot retrieval we are interested in the following research questions

- How effective is the language modelling approach compared to standard visual matching models such as Manhattan distance, Euclidean distance and Jensen-Shannon distance?
- What is the effect of different smoothing techniques on these low-level features? Are discounting methods as good as interpolation smoothing methods for these low-semantic features? Is one type of smoothing method more appropriate for some visual features or is there overall a single best smoothing model for all low-level features?
- What is the effectiveness of different visual features (colour, edge and texture) for video shot retrieval? Are regional features always better than global features? What discrete feature representation (number of dimensions and number of quantisation levels) is best for the different visual feature languages?

In our investigation of standard fusion methods for the discrete language modelling approach to video retrieval, we are interested in learning

- How to effectively combine language models of multiple visual features, multiple visual examples and multimodal features for video retrieval?

Our research is significant because it is the first to try to thoroughly investigate and compare different types of discrete language retrieval models on different visual features for the video search task In answering our research questions we are forwarding the knowledge in the research field on how best to match and fuse visual features for the video retrieval task We establish a baseline performance for many types of retrieval models using different features that other researchers may find useful when performing

their own studies We provide a viable discrete language modelling approach to video retrieval that achieves state-of-the-art performance as measured with the standard video retrieval test sets Our research may be further extended by other researchers such as in terms of better fusion strategies, relevance feedback and the transfer of other language modelling information retrieval techniques to the video retrieval task

1.4 Thesis Organisation

This thesis is organised into four main parts introduction, research proposal, evaluation and summary

Introduction

Chapter 1 In chapter 1 we provide a brief introduction to our research problem and outline the contents of this thesis

Chapter 2 In chapter 2 we introduce the language modelling approach to text information retrieval We present the overall structure of the language modelling approach and discuss language modelling approaches in terms of their probabilistic models, ranking models and smoothing techniques We also describe competing traditional retrieval models such as vector space models and probabilistic models

Chapter 3 In chapter 3 we describe the state-of-the-art in video retrieval, which involves matching based on text, visual and audio features and fusing these results We describe colour, texture, shape, motion and other spatio-temporal visual features that are popular in current video retrieval systems or otherwise promoted by the MPEG7 Multimedia Content Description Interface standard We also describe the matching models that are typically used to compare visual features We establish that many of the effective visual features, though not all, can be interpreted as languages of discrete symbols and therefore are amenable to retrieval using the discrete language modelling approach to information retrieval We describe existing comparative studies of video retrieval models and highlight their limitations in applicability to the video retrieval problem We finally present the TRECVid video retrieval benchmarking initiative, which provides the infrastructure for our empirical study of discrete language modelling approaches to the video retrieval problem

Research Proposal

Chapter 4 In chapter 4 we propose our video retrieval approach that uses discrete language models for visual retrieval and which is applicable to many of the effective visual features described in the previous chapter We propose three visual languages in our study representing the colour, edge and texture characteristics of video shots, which we use in much the same way as if they were text documents within the text-based language modelling information retrieval approach Due to the low-level nature of visual features, we propose the investigation of different statistical estimation smoothing techniques for these visual languages We outline extensions to the existing hierarchical language modelling approach for text-based video retrieval that include alternative semantic structures and alternative hierarchical smoothing techniques We also propose to evaluate standard score and rank-based fusion methods for combining these language models for video retrieval tasks

Evaluation In chapters 5, 6 and 7 we evaluate in sequence text-based language modelling approaches to video shot retrieval, our visual-based language modelling approach to video shot retrieval and the fusion of text and visual approaches

Chapter 5 In chapter 5 we evaluate the family of language modelling approaches for video shot retrieval using the TRECVid evaluation framework We compare the language modelling approaches with a set of representative text retrieval models We also compare different physical and semantic hierarchical structures and our alternative hierarchical smoothing techniques

Chapter 6 In chapter 6 we evaluate language models for different compact global representations of our colour, edge and texture features for video retrieval After establishing the relative performance of discrete language models and standard visual matching models on these visual features, we experiment with larger visual languages for grid-based regional variations of these features that take into account spatial location information We also evaluate the different physical and semantic hierarchical structures that we applied to the text feature in the previous chapter

Chapter 7 In chapter 7 we investigate different standard score and rank-based fusion methods for combining the text and visual retrieval language models We compare these efficient fusion strategies for combining visual features, combining the results of multiple visual examples and for combining multimodal features

Conclusions and Summary

Chapter 8 Finally, in chapter 8 we summarise our results, suggest extensions to our approach and describe future work

CHAPTER II

INFORMATION RETRIEVAL USING LANGUAGE MODELS

Language models are generative probability distributions for text sequences that use intricate smoothing methods to improve the quality of their estimates from sparse text samples At its most basic, the language modelling approach to information retrieval (LMIR) represents each document with its own language model and ranks documents based on their probability of generating the query This query-likelihood ranking is a simple and effective approach to information retrieval but lacks support for relevance feedback An alternative ranking based on the relative entropy between the query's language model and each document's language model provides a better mechanism for incorporating relevance feedback, but does not represent as formal an inclusion of relevance feedback as in the classical probabilistic IR models The main contribution of the LMIR approach is the intricate statistical estimation techniques and the simple generative framework, which can as easily be applied to information retrieval in other media such as image, video and audio as it is applied to text retrieval

2 1 Introduction

2 2 A Simple Language Model for IR

2 3 Probability Models

2 4 Statistical Estimation

2 5 Ranking Documents

2 6 Relevance Feedback

2 7 Non-Language Modelling Approaches to IR

2 8 Summary

2.1 Introduction

A language model is a generative probability distribution for text that models the probability of a sequence of words and can also directly generate sample text, though the generated text would be quite difficult to read and comprehend Language models were originally developed for the speech recognition task where they can improve the recognition rate and also reduce the search space (Jelinek, 1998) and have similarly been used for Optical Character Recognition, Machine Translation and other statistical work on text

Ponte and Croft (1998) proposed the language modelling approach to information retrieval (LMIR) where each document is represented by a language model and the documents are ranked by the query-likelihood – the probability that the document's

language model generates the query The language modelling approach to IR represents a new approach that is distinct from the classical IR models such as the Boolean model, the vector space model (Salton and Buckley, 1988) and the probabilistic models (Robertson and Sparck Jones, 1976, Crestani et al , 1998) LMIR primarily concerns text retrieval but can also be applied to multimedia retrieval since language models can represent image, video and audio features We will expand on this idea in later chapters

In LMIR the language model's parameters are first estimated from sample text such as a document's text and afterwards the language model is used for estimating probabilities of text samples such as a query's text The estimation of language model parameters requires careful attention especially for low frequency and missing words in the training text – otherwise, the probabilities of text sequences that contain these words will be given unreliable probabilities For example, a naive approach would give text sequences that contain words that were absent from the training text a zero probability To address these problems, the language modelling community has researched a wide range of methods for improving language model estimation that are collectively referred to as smoothing methods

In contrast to other applications of language models, information retrieval generally uses lower order n-grams such as unigrams where the probability of a word occurrence is independent of previous words in the text sequence The language modelling approach to information retrieval generally represents each document with its own language model, which also differs from traditional uses of language models where the tendency is to concentrate on a general model of the use of language The importance of smoothing in language models has also not diminished with their adoption to information retrieval and is actually essential in order to achieve effective retrieval due to the relatively small sample of text each document possesses

The rest of this chapter is organised as follows we first present a simple language modelling approach for information retrieval (Section 2 2), followed by a discussion of different probability models (Section 2 3), smoothing methods (Section 2 4), ranking methods (Section 2 5) and relevance feedback methods (Section 2 6) for language modelling approaches to information retrieval We compare LMIR with traditional approaches to IR such as the Boolean model, vector space model and probabilistic models in Section 2 7 and end the chapter with a brief summary

2.2 A Simple Language Model for IR

In this section we will present a simple unigram language model for information retrieval A language model is a probability distribution for text sequences, which for a text

sequence, $w_1 \quad w_n$, can be expressed as

$$\begin{aligned}\mathbb{P}(w_1 \quad w_n) &= \prod_i^n \mathbb{P}(w_i|w_1 \quad w_{i-1}) \\ &= \mathbb{P}(w_1)\mathbb{P}(w_2|w_1)\mathbb{P}(w_3|w_1,w_2)\mathbb{P}(w_4|w_1\ w_2,w_3) \quad \mathbb{P}(w_n|w_1 \quad w_{n-1}),\end{aligned} \tag{1}$$

by using the chain rule expansion of joint probability Language models differ in their approximation of this by making different independence assumptions and by using different estimation strategies for the individual word probabilities

The unigram language model makes the most strict independence assumption and assumes that the probability of a word in a sequence does not depend on any of the previous words The unigram multinomial language model for the probability of a text sequence, which loosely approximates Equation (1), is defined as

$$\mathbb{P}(w_1 \quad w_n) = \prod_i^n \mathbb{P}(w_i) = \mathbb{P}(w_1)\mathbb{P}(w_2)\mathbb{P}(w_3)\mathbb{P}(w_4) \quad \mathbb{P}(w_n) \tag{2}$$

The unigram multinomial language model was first proposed for language modelling-based information retrieval in (Hiemstra, 1998) and is popular in other LMIR approaches (Song and Croft, 1999, Berger and Lafferty, 1999a, Lafferty and Zhai, 2001) but is less useful than bigram and trigram approximations of Equation (1) for the traditional applications of language models such as speech recognition

The query-likelihood for the multinomial unigram model is the probability of drawing *in sequence* the query terms from the document's multinomial unigram distribution, which is defined as

$$\mathbb{P}(\mathbf{q}|\mathbf{M_d}) = \prod_{i=0}^{|q|} \mathbb{P}(q_i|\mathbf{M_d}), \tag{3}$$

where $\mathbf{q}$ is the sequence of query terms and $\mathbf{M_d}$ is the document's language model The parameters of the document's unigram language model are simply all the probabilities of the individual terms and are estimated from the document text A direct estimate of the probability of a term using the maximum likelihood estimate (relative frequency of the term in the document) is ineffective since it gives a probability of zero to all absent terms Zero probability is a very severe estimate for any term as it means the language model can never generate that term or any sequence of words containing it This problem is exacerbated as a document is a very small and sparse text sample compared to the complete vocabulary If we use language models estimated in this way in a query-likelihood retrieval system, then all documents that are missing any or all query terms will be given the *same probability of zero* for generating the query

The standard methods for addressing this problem in language modelling are collectively called smoothing methods (see Section 2 4), which redistribute some of the probability that is normally given to terms observed in the document to missing and

low frequency terms The smoothing methods used in LMIR often involve combining the term's maximum likelihood estimate with a background estimate based on the whole collection, as originally proposed in (Ponte and Croft, 1998) A straightforward and popular method for combining two probability estimators is a finite mixture model, a linear interpolation with weights that sum to one, which was first used for LMIR in (Hiemstra, 1998) The linear interpolated probability of a unigram term t in document d using the local document term probability and the collection probability, which is often referred to as Jelinek-Mercer smoothing, is defined as

$$\mathbb{P}_{JM}(t|M_d) = (1-\lambda)\mathbb{P}_{ML}(t|d) + \lambda\mathbb{P}_{ML}(t), \tag{4}$$

where $\mathbb{P}_{ML}(t|d)$ is calculated as

$$\mathbb{P}_{ML}(t|d) = \frac{tf(d,t)}{dl_d} \tag{5}$$

and the background probability of a term in the collection $\mathbb{P}_{ML}(t)$ is calculated by relative collection frequency, that is

$$\mathbb{P}_{ML}(t) = \frac{cf(t)}{cs} \tag{6}$$

In these equations $tf(d,t)$ is the frequency of the term in the document, dl_d is the document length, $cf(t)$ is the number of times term t is present in the collection and cs is the number of terms in the collection An alternative is to calculate the background probability using document frequency (Hiemstra, 1998) but this reduces the amount of information used in the background estimate

The basic unigram language model for information retrieval that combines the local document's estimate with the collection term estimate in a finite mixture model is now completely defined The only free variable in this retrieval model is λ, which controls the amount of smoothing with the background collection probabilities This can be set empirically by testing on a sample collection

The query-likelihood retrieval status value (RSV) for each document, which is used to rank documents, can be directly calculated by combining equations (4), (5), and (6), as

$$RSV_{d,q} = \mathbb{P}(q|d) = \prod_{i=0}^{|q|}\left(\lambda\frac{cf(q_i)}{cs} + (1-\lambda)\frac{tf(q_i,d)}{dl_d}\right), \tag{7}$$

which is no more expensive to evaluate than the TF-IDF model since the query-likelihood RSV can be expressed in term of the unique query terms that are present in the document

$$RSV_{q,d} = \mathbb{P}(q|d) = \prod_{i=1}^{|q|}\lambda\frac{cf(q_i)}{cs} \times \prod_{q_i\in d}\left(\left(\frac{cs}{cf(q_i)}\right)\left(\lambda\frac{cf(q_i)}{cs} + (1-\lambda)\frac{tf(q_i,d)}{dl_d}\right)\right)^{qf(t_i)} \tag{8}$$

$$\propto \prod_{q_i\in d}\left(\left(\frac{cs}{cf(q_i)}\right)\left(\lambda\frac{cf(q_i)}{cs} + (1-\lambda)\frac{tf(q_i,d)}{dl_d}\right)\right)^{qf(t_i)} \tag{9}$$

It is straightforward to extend these formulae for bigram and higher order n-gram models and we could smooth these higher order models by interpolating them with the lower-order n-gram models However, while it is common to use the bigram and trigram language models in speech recognition, their use in LMIR is problematic Unlike speech recognition, in LMIR a language model is estimated for each document and the document's size would particularly not support the accurate estimation of the trigram-based language models Also, the use of higher order n-grams would negatively impact the query-likelihood estimation as queries are phrased differently to documents and have a different composition style Bigrams are used in LMIR to achieve some level of phrase searching (Miller et al , 1999a, Song and Croft, 1999) but in general current language modelling approaches to information retrieval predominantly use unigram language models

2.3 Probability Models

The original language modelling approach to information retrieval represents queries as being sampled from a Multiple-Bernoulli distribution that is estimated for each document (Ponte and Croft, 1998) Queries are modelled as a set of unigram terms and the documents are ranked by query-likelihood, which for the Multiple-Bernoulli language model is defined as

$$\mathbb{P}(\mathbf{q}|\mathbf{M_d}) = \prod_{t \in \mathbf{q}} \mathbb{P}(t|\mathbf{M_d}) \times \prod_{t \notin \mathbf{q}} 1 - \mathbb{P}(t|\mathbf{M_d}), \tag{10}$$

where $\mathbf{q}$ is the set of unigram query terms, $\mathbf{M_d}$ is the document's language model, and t is any unigram term from the entire vocabulary The significance of this work is in placing information retrieval into a language modelling framework where documents are ranked by query-likelihood However, the Multiple-Bernoulli set-based representation of documents and queries complicates the estimation of the model's parameters for a specific document and restricts the representation of the query to binary weighted terms

In contrast, in the current language modelling approaches to information retrieval a multinomial distribution is used to model documents and queries (Hiemstra, 1998, Song and Croft, 1999, Berger and Lafferty, 1999a, Lafferty and Zhai, 2001), which defines query-likelihood as

$$\mathbb{P}(\mathbf{q}|\mathbf{M_d}) = \prod_{i=1}^{|q|} \mathbb{P}(q_i|\mathbf{M_d}) \tag{11}$$

where $\mathbf{q}$ is a sequence of query terms and $\mathbf{M_d}$ is the document model In this model queries and documents are modelled as a sequence of terms sampled individually from a multinomial distribution The benefit of this probability model is that queries can have multiple occurrences of the same term and the language model's parameters can be more directly estimated from the document's text than when using the Multiple-Bernoulli model

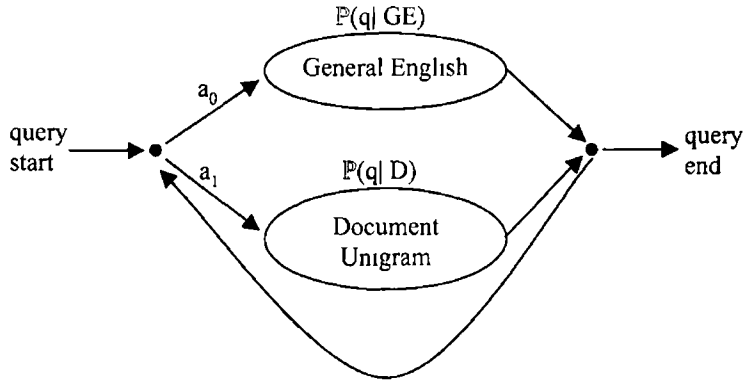


Figure 1 Language modelling approach to information retrieval using query-likelihood from a simple 2-state hidden Markov model of each document (Miller et al , 1999a,b)

An equivalent probability model to the multinomial language model was introduced as an hidden Markov model (HMM) approach to information retrieval in (Miller et al , 1999a,b) Their proposed 2-state unigram HMM model that combines the document model with the general English model is essentially a multinomial unigram model with interpolated (Jelinek-Mercer) smoothing from an HMM perspective (see Figure 1)

2.4 Statistical Estimation

The statistical estimation of the language model's parameters from sample text is well studied in the statistical language modelling community The statistical estimation techniques used in language modelling are referred to as smoothing and are rooted in the firm foundations of the general field of statistical estimation and inference The earliest smoothing technique, Laplace's Law (translated in Laplace (1995)), predates the statistical language models by over a century

The unsmoothed estimate is the maximum likelihood estimate The concept of maximum likelihood estimation (Fisher, 1922) of model parameters is the selection of the parameter values of a probability model so as to give maximum probability to the training sample The maximum likelihood estimate for an unigram w is its relative frequency in the sample text and is defined as

$$P_{ML}(w) = \frac{C(w)}{N}, \tag{12}$$

where N is the number of unigrams in the sample text and $C(\,)$ is the number of occurrences of the given unigram

The smoothed estimate tries to better handle unobserved and low frequency n-grams and there are two general approaches to smoothing – the first is to adjust the maximum likelihood estimate to better take into account unobserved and low-frequency n-grams

treating all unobserved events similarly and the second general approach is to combine different probability estimators together

Smoothing is used in traditional language modelling in order to allow higher order n-grams to be used in the language model and most of the following smoothing techniques can be found in a standard text on statistical natural language processing such as (Manning and Schutze, 1999) In the language modelling approach to IR, the document's language model requires significant attention to smoothing even when based on unigram probabilities Even before using higher order n-grams, the unigram model will produce unreliable estimates for missing terms if the maximum likelihood estimate is used without smoothing

In the following subsections we will describe the different smoothing methods for unigram language models grouped into the following categories discounting, validation, interpolation and back-off models We will describe these smoothing methods for the unigram document language model using the following notation

- $\mathbb{P}_D(w)$ the unsmoothed estimate for the term w in the document
- $\mathbb{P}_C(w)$ the unsmoothed estimate for the term w in the collection
- $C(w)$ count of term w in document sample
- N size of document sample
- N_r number of terms with r frequency in document sample (e g N_0 is the number of terms that have zero frequency)
- V number of unique term in the document sample
- B size of vocabulary

2 4 1 Discounting-Based Smoothing

Discounting techniques treat all unobserved terms equally by adding a small probability to unobserved terms and normalising the probability mass to sum to 1 The discounting methods differ in how they redistribute the probability mass and are generally unsuitable for smoothing text language models in LMIR for the fact that they treat all missing words exactly the same However, they represent the classical statistical approach to smoothing, and can be applied, if necessary, to smooth the background collection probability estimates They may also be useful for smoothing visual languages that have a more uniform distribution and that are less semantic than text languages

Laplace smoothing (Laplace, 1995) Laplace smoothing adds one to each term's frequency and is defined as

$$\mathbb{P}_{lap}(w) = \frac{C(w) + 1}{N + B} \tag{13}$$

This essentially is adding one event for every type of event we are modelling in the distribution Unfortunately, this gives too much probability to unobserved events and the larger the size of the vocabulary the more probability that is taken away from observed terms In text IR, documents are very small samples, and this "adding one" rule overwhelms the probability distribution with the amount of probability given to the unobserved terms Even for traditional applications of LM this method is widely criticised (Gale and Church, 1994)

Lidstone smoothing (Hardy, 1889, Lidstone, 1920) Lidstone smoothing is a simple modification of Laplace smoothing that adds a small value λ to each count and is defined as

$$\mathbb{P}_{lid}(w) = \frac{C(w) + \lambda}{N + B\lambda}, \tag{14}$$

where λ is a parameter for our prior belief in the uniformity of the events over the empirical observations The parameter setting of $\lambda = 1/2$ is a common choice, which is referred to as the Jeffrey's prior

Both Laplace and Lidstone smoothing are special cases of the more general Dirichlet smoothing (Bayesian smoothing) that assume that the prior distribution is uniform We can interpret these smoothing methods as adding either 1 or λ virtual counts respectively to each term of the empirical distribution A nice property of these discounting methods is that the larger the sample size the less probability is redistributed This is intuitive as a larger document should be more representative of its topic and therefore require less smoothing

Absolute Discounting (Ney et al , 1994) Absolute Discounting subtracts a small constant δ from the frequency of observed terms and evenly distributes this freed frequency to unobserved terms Absolute discounting is defined as

$$\mathbb{P}_{abs}(w) = \begin{cases} \frac{C(w)-\delta}{N} & \text{if } C(w) > 0, \\ \frac{(B-N_0)\delta}{N_0 N} & \text{otherwise,} \end{cases} \tag{15}$$

where the parameter δ controls the amount of smoothing Similar to Lidstone the larger the document the less smoothing but also the less zero frequency terms the more smoothing, which could provide quite erratic smoothing in documents with near full coverage of their language This is unlikely for text documents but can occur for documents using visual languages that have a small vocabulary

Linear Discounting (Ney et al , 1994) In Linear Discounting, the maximum likelihood estimates for non-zero frequency events are scaled with the freed up mass allocated to zero frequency events Therefore, a fixed proportion of the probability mass, α, is distributed evenly to each of the unobserved events The smaller the number of unobserved events the more probability each unobserved event individually gets, which is counter-intuitive for the information retrieval task Linear discounting is defined as

$$\mathbb{P}_{lin}(w) = \begin{cases} (1-\alpha)\frac{C(w)}{N} & \text{if } C(w) > 0, \\ \frac{\alpha}{N_0} & \text{otherwise,} \end{cases} \tag{16}$$

where parameter α is $0 \leq \alpha \leq 1$

Good-Turing Estimation (Good, 1953) In Good-Turing estimation we adjust the frequencies based on the assumption that the underlying distribution is binomial The adjusted frequency $r*$ is given by

$$r* = (r+1)\frac{\mathbb{E}(N_{r+1})}{\mathbb{E}(N_r)}, \tag{17}$$

where $\mathbb{E}(N_r)$ is the expected number of words that occur r times The empirical estimate of $\mathbb{E}(N_r)$ is not a good estimate when r is large and to improve this estimate a function $S(N_r)$ can be fit using statistical regression (Manning and Schutze, 1999), producing the following Good-Turing estimate

$$\mathbb{P}_{GT}(w_1 \quad w_n) = \begin{cases} \frac{r*}{N}, \quad \text{where } r* = (r+1)\frac{S(r+1)}{S(r)} & \text{if } r > 0 \\ \frac{1-\sum_{r=1}^{\infty} N_r \frac{r*}{N}}{N_0} \approx \frac{N_1}{N_0 N} & \text{if } r = 0 \end{cases} \tag{18}$$

The Good-Turing estimate has been successfully used in traditional language modelling applications but for LMIR there is not enough frequency information to use it reliably in document modelling However, the Good-Turing estimate has been applied to IR language models and evaluated for ad hoc retrieval in (Song and Croft, 1999)

2 4 2 Combination-Based Smoothing

Discounting methods treat unobserved words equally, while a more effective smoothing method in text LMIR is to combine the document's maximum likelihood estimate (MLE) with a background model of English such as the collection model, which allows terms to be smoothed in proportion to how often they naturally occur The combination-based smoothing methods in this section, Jelinek-Mercer smoothing, Dirichlet smoothing, and Absolute interpolation were evaluated for ad hoc text retrieval in (Zhai and Lafferty, 2001b) Their results were not conclusive but suggest that Dirichlet smoothing is best for short queries and either Jelinek-Mercer or Dirichlet are best for long queries They interpreted this as implying that Dirichlet smoothing could be better at estimating document models, whereas Jelinek-Mercer smoothing may be good at modelling non-informative words in the query

Jelinek-Mercer smoothing (Zhai and Lafferty, 2001b) Jelinek-Mercer smoothing is a simple finite mixture model (linear interpolation) of the document's MLE with a background distribution (collection model) and is defined as

$$\mathbb{P}_{JM}(w) = (1 - \lambda)\mathbb{P}_D(w) + \lambda\mathbb{P}_C(w), \tag{19}$$

where a suitable value for the parameter λ must be chosen In contrast to other combination-based smoothing methods, this smoothing method keeps the influence of the background model constant for each document

Dirichlet smoothing (Zhai and Lafferty, 2001b) In Dirichlet smoothing (empirical Bayes smoothing) the document's MLE is mixed with a background distribution, similar to Jelinek-Mercer smoothing, but also taking into account the size of the document by reducing the effect of the background distribution for larger documents The smaller the document the more smoothing required Dirichlet smoothing is defined as

$$\mathbb{P}_{DIR}(w) = \frac{N}{(N+\mu)}\mathbb{P}_D(w) + \frac{\mu}{(N+\mu)}\mathbb{P}_C(w), \tag{20}$$

where the parameter μ controls the amount of smoothing and represents the strength of our prior belief in the background probability over the document's empirical distribution This method is similar to the Lidstone discounting method except that the prior distribution is no longer assumed to be uniform The μ parameter can be interpreted as the number of virtual terms taken from the collection distribution and added to the empirical document distribution

Witten-Bell Smoothing (Witten and Bell, 1991, Lavrenko, 2000) Witten-Bell smoothing combines the document's MLE with the collection's probability model and adjusts the mixing parameter depending on how much redundancy there is in the document The more redundancy, the less smoothing is required Redundancy is related to V the number of unique terms in the document This method does not require parameter tuning and is defined as

$$\mathbb{P}_{WB}(w) = \frac{N}{(N+V)}\mathbb{P}_D(w) + \frac{V}{(N+V)}\mathbb{P}_C(w) \tag{21}$$

Absolute interpolation smoothing (Zhai and Lafferty, 2001b, Ney et al , 1994) Absolute interpolation, an extension of standard absolute discounting, subtracts a constant amount from observed terms and this freed probability mass is redistributed using the background probability distribution Absolute interpolation is defined as

$$\mathbb{P}_{ABS}(w) = \begin{cases} \frac{C(w)-\delta}{N} + \frac{(B-N_0)\delta}{N}\mathbb{P}_C(w) & \text{if } C(w) > 0 \\ \frac{(B-N_0)\delta}{N}\mathbb{P}_C(w) & \text{otherwise} \end{cases}, \tag{22}$$

where the parameter δ controls the amount of frequency taken from observed events Similar to Absolute discounting, the larger the sample size the less smoothing and the more terms that have non-zero frequency the more smoothing However, this smoothing applies to all terms in Absolute interpolation and therefore is probably more stable than Absolute discounting for documents with a relatively small number of zero frequency terms

2 4 3 Back-Off Smoothing

Back-off models are an alternative to interpolation methods for combining estimators When estimators are combined in back-off models they are ordered by their specificity and the most specific model is used if it is suitable, otherwise, the next most specific is used, and so on, until a suitable model is found Interpolation-based smoothing has been reported to be consistently superior to back-off estimators for ad hoc searching (Zhai and Lafferty, 2001b) and so we will not consider them further

2 4 4 Validation-based Smoothing

In validation smoothing methods, we divide the training sample into smaller samples We can then train on one part of the sample and choose the amount of smoothing by validating the model on the other part The importance of these methods is that they can optimise the selection of the control parameters of another smoothing method

Leave-One-Out In the Leave-One-Out validation method, each word is left out in turn to create N simulated tests and the amount of smoothing is chosen to maximise the overall likelihood of the held out token over all N tests The Leave-One-Out method is a validation method that has been used for LMIR to automatically control the value of the parameter for the Dirichlet smoothed document language models (Zhai and Lafferty, 2002)

2.5 Ranking Documents

In the standard language modelling approach documents are ranked based on query-likelihood but in more recent language modelling approaches that perform relevance feedback the documents are ranked using relative entropy

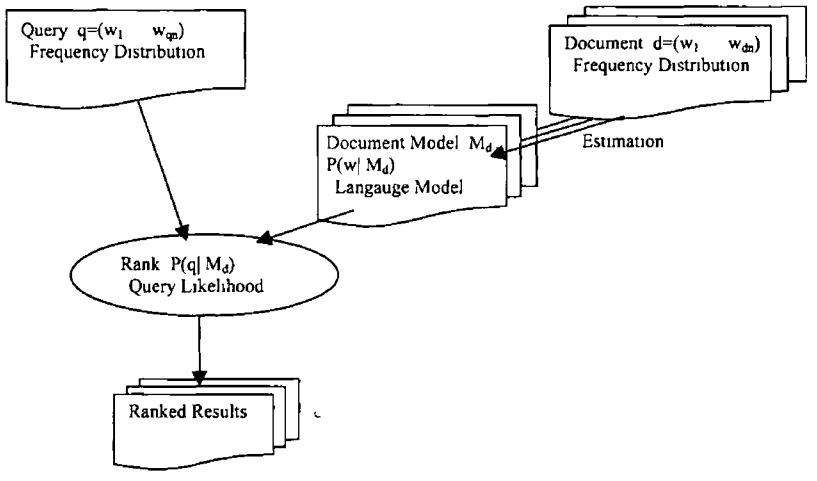


Figure 2 Language modelling approach to information retrieval using query-likelihood ranking

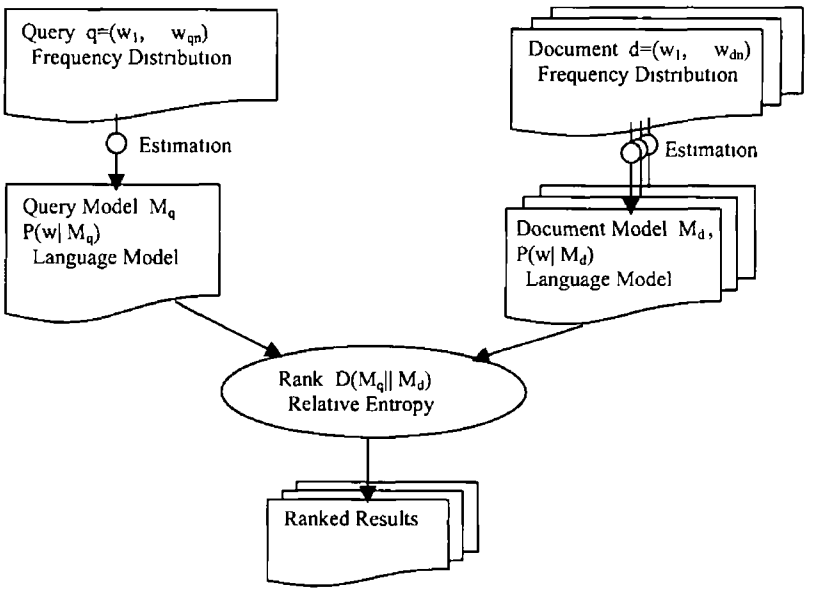


Figure 3 Language modelling approach to information retrieval using relative entropy (KL divergence, cross entropy) ranking

2.5.1 Query-Likelihood

Query-likelihood ranks documents based on the probability that they would generate the query. The advantage of this ranking over document-likelihood is that the documents are larger samples of text and therefore provide more reliable probability estimates than the generally smaller query text. The query-likelihood method compares the probability of the same set of events for all documents, which ensures documents are sampled for all terms in the query and this leads to a retrieval status value (RSV) that represents a more exhaustive match between the query and documents than document-likelihood. The query-likelihood ranking also provides an explanation for the language modelling approach that describes the process of searching as a user choosing query terms based on their likelihood of being present in an ideal document that would satisfy their information need. By applying query-likelihood we are representing this hypothetical process and trying to find the ideal document, which we presuppose is related to the relevance of the document to the information need. The primary problem with query-likelihood is that it removes the concept of relevance from the retrieval model and more importantly it does not support a natural way of including relevance feedback.

Zaragoza et al. (2003) propose a Bayesian extension to the query-likelihood language modelling approach that uses a full Bayesian predictive distribution when calculating the query-likelihood. This approach is very similar to the Dirichlet smoothed language model in that it combines the document sample's likelihood with the prior collection distribution, but goes one step further by modelling the uncertainty of the resulting parameters and integrating this uncertainty out when using the model to predict the query. The Bayesian query-likelihood or posterior predictive distribution for the multinomial distribution with Dirichlet prior is defined as

$$P(q|d,n_\alpha) = \frac{\Gamma(n_q+n_\alpha)}{\prod_{i=1}^{V}\Gamma(q_i+\alpha_i)}\frac{\Gamma(n_{d_l}+n_\alpha)}{\prod_{i=1}^{V}\Gamma(d_i+\alpha_i)}\int\prod_{i=1}^{V}\theta_i^{q_i+d_{l,i}+\alpha_i-1}, \tag{23}$$

where n_α is a parameter specifying the strength of our belief in the prior distribution and $\alpha_i = n_{\alpha_i}\mathbb{P}(w_i|C)$ is the prior count for each term when using the collection model as the prior distribution. The RSV for this query-likelihood ranking can be more efficiently calculated (see Zaragoza et al. (2003) for details) as

$$RSV_{q,d} = log(q|d_l) \propto \sum_{i|(q_i,d_i)\neq 0}\sum_{g=1}^{q_i}\log\left(1+\frac{d_{l,i}}{\alpha_i+g-1}\right) - \sum_{j=1}^{n_q}\log(n_{d_l}+n_\alpha+j-1) \tag{24}$$

This Bayesian query-likelihood is slightly less efficient than the standard query-likelihood as the number of operations to compute it is related to the number of matching terms as opposed to the number of matching term indexes. This is not a significant problem for ad hoc text retrieval but causes speed problems when used on visual languages that have large query samples to predict such as over 50,000 matching terms but

are from relative small visual languages and therefore are quite efficient for the standard query-likelihood function

2 5 2 Document-Likelihood

Document-likelihood is the probability of the query generating the document and is defined as

$$\mathbb{P}(D|Q) = \prod_{t \in D} \mathbb{P}(t|Q)^{c(t,D)} \tag{25}$$

where $c(t, D)$ is the term count of term t in document D The problem with the document-likelihood for a multinomial language model is that longer documents are less likely than shorter documents and common words will dominate To correct this severe bias the documents can be ranked by the likelihood ratio

$$\frac{\mathbb{P}(D|Q)}{\mathbb{P}(D|C)} = \prod_{t \in D} \left(\frac{\mathbb{P}(t|Q)}{\mathbb{P}(t|C)} \right)^{c(t,D)} \tag{26}$$

where $\mathbb{P}(t|C)$ is the probability of a term in the background collection The problem with this approach is that the predicted terms are different for each document and therefore in some way this probability is not comparable and also the document-likelihood may only represent matching between subparts of the query Since the query is generally very small, document-likelihood estimates based on it are highly unreliable compared to the query-likelihood, which bases its probability estimation on the generally larger sample of document text The ideal document according to the document-likelihood ratio is the document with the highest ratio repeated for its whole document, which could not be considered a desirable outcome These problems were also observed when using document-likelihood with relevance models to rank documents (see Section 2 6 2 for details on relevance models), which used the following ratio

$$\frac{\mathbb{P}(D|R)}{\mathbb{P}(D|\bar{R})} = \prod_{t \in D} \left(\frac{\mathbb{P}(t|R)}{\mathbb{P}(t|C)} \right)^{c(t,D)} \tag{27}$$

where R is the relevance language model Lavrenko and Croft (2003) noted that this ratio suited documents that repeated the term with maximum likelihood ratio and found that cross entropy was a better method for comparing relevance models with document models

2 5 3 Relative Entropy and Cross Entropy

A more recent language modelling approach (Lafferty and Zhai, 2001), represents both queries and documents using language models and ranks documents based on the relative entropy between the query's language model and the document's language model (see

Figure 3), which allows for relevance feedback to be more easily incorporated into the query representation

Relative entropy $D(X\|Y)$, also called Kullback-Leibler divergence/distance, is an information theoretic measure of the difference between two probability distributions over the same event space It measures the amount of information that is needed to encode events from the first probability distribution into the second It evaluates to a non-negative quantity that is zero only when both probability distributions are the same It is not strictly a distance function as it does not satisfy the triangle inequality and is not symmetric The relative entropy measure was first used for LMIR ranking in (Lafferty and Zhai, 2001) where the query is represented as a probability distribution that takes into account translation probabilities defined by a random-walk Markov chain technique To use relative entropy for ranking we represent both the query and the documents using language model The relative entropy between the query's language models, $\mathbf{M_q}$, and the document's language model, $\mathbf{M_d}$, is defined as

$$RSV_{q,d} = D(\mathbf{M_q}\|\mathbf{M_d}) = \sum_w P(w|\mathbf{M_q}) \log \frac{P(w|\mathbf{M_q})}{P(w|\mathbf{M_d})} \tag{28}$$

The RSV for ranking documents using relative entropy can be simplified, as the query is constant for ranking a set of documents, giving

$$RSV_{q,d} = -\sum_w P(w|\mathbf{M_q}) \log P(w|\mathbf{M_d}), \tag{29}$$

which is actually the formula for a common measure in statistical language modelling called cross entropy Furthermore, if the query's language model is calculated by maximum likelihood with no smoothing, then the scoring function is simply a re-scaled version of the query's log-likelihood, which of course produces the same ranking as query-likelihood

2.6 Relevance Feedback

Language modelling-based relevance feedback is limited when using query-likelihood as it constrains the query representation to a frequency distribution of terms The relative entropy ranking method provides a more consistent and direct approach to relevance feedback as it supports updating the query's language model and the comparison of relevance-based language models with documents In the following subsections we will describe some approaches to language modelling-based relevance feedback We will denote the set of positive feedback documents as $\mathcal{F}$, which may simply be pseudo-relevant documents such as a fixed number of top ranking documents from the initial search results

2 6 1 Query-Likelihood Feedback

Ponte (1998) supports limited relevance feedback in the Multiple-Bernoulli language model by adding terms to the query based on the ratio of their probability in feedback documents compared to their probability in the background collection The score, $s(w)$, for each word is defined as

$$s(w) = \prod_{d \in \mathcal{F}} \frac{P(w|\mathbf{M_d})}{\mathbb{P}(w|C)} \tag{30}$$

This ad hoc feedback method is limited as it only supports binary weighted queries

In the HMM retrieval model (Miller et al , 1999a) transition probabilities are trained in order to perform relevance feedback based on a term's frequency in relevant documents and the collection As noted by Zhai and Lafferty (2001a), the transition probabilities are estimated in an heuristic manner and the HMM model is no longer equivalent to the language modelling approach

Hiemstra (2002) equates smoothing with the importance of a query term and proposes a term-specific smoothing method The term-specific smoothing method extends the Jelinek-Mercer smoothed unigram language model with query specific smoothing parameters and is defined as

$$P(\mathbf{q}|\mathbf{M_d}, \lambda) = \prod_{i=1}^{|q|} (\lambda_i \mathbb{P}(q_i|\mathbf{M_d}) + (1 - \lambda_i)(P(q_i|C)), \tag{31}$$

where the λ_i term importance weights (query term smoothing parameters) are estimated from the relevant documents using the EM algorithm (Dempster et al , 1977) in order to optimise the probability of the query terms given the relevant documents This limited method of feedback does not expand the query but provides a re-weighting of the query's terms and it complicates the notion of parameter estimation for the document model by equating it with query-specific smoothing of search terms

The unigram language model was extended to take into account term translation probabilities, thereby modelling information retrieval as machine translation (Berger and Lafferty, 1999a,b) This approach enabled the retrieval of documents that do not contain the query terms by allowing synonyms and other relationships between words to be taken into account in their model This model is essentially a language modelling approach that supports query expansion but does not provide any wider relevance feedback functionality As recognised in (Lafferty and Zhai, 2001) this approach is inefficient as it requires a sum over all terms in the document and is also hindered by the large amount of training data required to perform query specific translation

2 6 2 Query Model Feedback

Ranking documents with the information theoretic measure relative entropy or equivalently cross entropy defines the language modelling approach in terms of the difference in information between the query's language model and each document's language models The representation of the query as a probability distribution provides a more powerful mechanism for supporting relevance feedback because it allows more intricate probabilistic modelling and updating of the query representation than possible in the query-likelihood formulation

Lafferty and Zhai (2001) first use relative entropy for ranking documents by the information theoretic comparison of the query's and documents' language models and introduce a Markov chain approach to estimate the query model that can be used for query expansion and pseudo-relevance feedback by re-estimating the initial query model

A more direct relevance feedback language modelling approach that supports query model updating is proposed in (Zhai and Lafferty, 2001a), which suggests two approaches for generating a feedback model based on a set of relevant (or pseudo-relevant) documents In their first feedback model approach, referred to as a generative model of feedback documents, they assume that a feedback unigram model generated the relevant documents and estimate its parameters using the EM algorithm, which maximises the following likelihood of the feedback documents

$$\mathbb{P}(\mathcal{F}|\theta_{\mathcal{F}}, \lambda) = \prod_{d \in \mathcal{F}} \prod_{w} ((1-\lambda)\mathbb{P}(w|\theta_{\mathcal{F}}) + \lambda \mathbb{P}(w|\mathcal{C}))^{c(w,d)} \tag{32}$$

The mixture parameter λ that controls the amount of noise from the collection model is set to a constant when the feedback model's parameters $\theta_{\mathcal{F}}$ are estimated using the EM algorithm This estimation is presented as a purification of the feedback model (topic model) by the elimination of the effects of the background noise (common words) Their second feedback model approach, referred to as divergence minimisation over feedback documents, chooses the query model that has the smallest average relative entropy between the feedback model and the smoothed feedback documents The motivation is to achieve the best average retrieval score (relative entropy) for the feedback documents They add a regularising term to the divergence minimisation feedback approach in order reduce the impact of common words from the collection, which leads to the following definition of their feedback model

$$\hat{\theta}_{\mathcal{F}} = \arg\min_{\theta} \frac{1}{|\mathcal{F}|} \sum_{d \in \mathcal{F}} D(\theta \,||\, \hat{\theta}_d) - \lambda D(\theta \,||\, \theta_C) \tag{33}$$

where the parameter λ controls the weight given to the collection model This results in a closed form expression for the calculation of the query model that unlike the first feedback method does not require the iterative EM procedure For both these methods of estimating the query model from feedback documents the updated query model is an

interpolation of the original query model and the feedback model and is defined as

$$\hat{\theta}_{Q'} = (1-\alpha)\hat{\theta}_Q + \alpha\theta_{\mathcal{F}} \tag{34}$$

where α controls the strength of the feedback model and is set empirically Both these methods improve on the non-feedback results and achieve slightly better results than Rocchio feedback (Rocchio, 1971) under similar test conditions It was also found that the first method, generative model of feedback documents, is more stable with respect to its parameters than the divergence minimisation method

A quite different approach to relevance feedback is the combination of relevance models $\mathbb{P}(w|R)$ with the language modelling approach to information retrieval (Lavrenko and Croft, 2001) that similar to classical probabilistic models ranks documents by their probability of relevance (Probability Ranking Principle, see Robertson (1977)) They equivalently express this criteria as the odds of the document being observed in the relevant class and model both relevant (R) and non-relevant (N) classes using unigram language models

$$RSV_{d,q} = \frac{\mathbb{P}(D|R)}{\mathbb{P}(D|N)} \approx \prod_{w \in D} \frac{\mathbb{P}(w|R)}{\mathbb{P}(w|N)} \approx \prod_{w \in D} \frac{\mathbb{P}(w|q_i \quad q_k)}{\mathbb{P}(w|C)} \tag{35}$$

They approximate $\mathbb{P}(w|N)$ using $\mathbb{P}(w|C)$ and argue that $\mathbb{P}(w|q_i \quad q_k)$ is a good approximation of $\mathbb{P}(w|R)$ when no relevance information is available In contrast to standard language modelling interpretations they assume that query and documents are samples from an unknown relevance model and propose two methods for estimating the relevance model when only the query and no relevance judgements are available The first method assumes identical and independent sampling of a query model to generate both the original query terms and the words in the pseudo-relevant documents, which leads to the following joint probability of a word and the existing query terms

$$\mathbb{P}(w, q_1 \quad q_k) = \sum_{d \in \mathcal{F}} \mathbb{P}(d)\mathbb{P}(w|\mathbf{M_d}) \prod_{j=1}^{k} \mathbb{P}(q_j|\mathbf{M_d}) \tag{36}$$

The relevance unigram model is simply estimated using the conditional probability, which is expressed in terms of this joint probability

$$\mathbb{P}(w|R) \sim \mathbb{P}(w|q_1 \quad q_k) = \frac{\mathbb{P}(w, q_1 \quad q_k)}{\mathbb{P}(q_1 \quad q_k)} \tag{37}$$

In the second method for estimating the relevance class they use conditional sampling which adds words to the query by first choosing the word w based on prior $\mathbb{P}(w)$, then independently for each existing query word selecting a pseudo-relevant document distribution based on $\mathbb{P}(\mathbf{M_d}|w)$ and then sampling the query word based on $\mathbb{P}(q_i|\mathbf{M_d})$, which leads to the following joint probability

$$\mathbb{P}(w, q_1 \quad q_k) = \mathbb{P}(w) \prod_{i=1}^{k} \sum_{d \in \mathcal{F}} \mathbb{P}(q_i|\mathbf{M_d})\mathbb{P}(\mathbf{M_d}|w) \tag{38}$$

Their approach to relevance feedback is not integrated with the initial querying and is quite separate to the initial search model First they retrieve documents using the Jelinek-Mercer language model, then take the top 50 documents and estimate the relevance model and finally use the odds of relevance to rank documents replacing completely their original retrieval model They found that their two estimation methods achieve statistically significantly better results than using the Jelinek-Mercer language model alone with no relevance feedback

Lavrenko and Croft (2003) compare this relevance ratio approach to ranking documents with the cross entropy of the relevance language model and each document's language model and found that the cross entropy ranking performs better than using the likelihood ratio They showed that the likelihood ratio of the relevance class when assuming term independence favours documents containing many occurrences of a few words with highest $\frac{\mathbb{P}(w|R)}{\mathbb{P}(w|N)}$, whereas the cross entropy approach compares all documents based on the same set of events within the relevance probability distribution and therefore achieves better results

2.7 Non-Language modelling approaches to Information Retrieval

In this section we discuss three different types of retrieval models, the Boolean model, the vector space model, and the probabilistic models, which fundamentally differ in how they model the information retrieval task

2 7 1 Boolean Model

The Boolean model is a set-based retrieval model that supports querying using multiple query terms combined with AND, OR and NOT logic Each term in the indexing language is mapped to a set of documents and the AND, OR and NOT Boolean operators are simply implemented as respectively intersection, union and complement of these sets of documents The Boolean model provides no ranking of retrieved documents and simply separates the documents into relevant and non-relevant sets, which limits its usefulness for real-world information retrieval tasks Another problem with the Boolean model is that users (even computer science graduates) find it hard to express an information need in Boolean logic, and simply AND-ing or OR-ing a set of query terms will likely give too few or too many results respectively

Coordinate Level Ranking is a somewhat related retrieval model that supports partial matching between documents and queries In Coordinate Level Ranking the documents that match more query terms are guaranteed to be ranked higher in the results list

than those matching less terms A problem with both the Boolean and Coordinate Level Ranking is that all query and document terms are treated equally ignoring their frequency of occurrence in the documents or the wider collection

The MLE language model, a language model without smoothing, has similar problems to a Boolean AND-ed set of query terms, since it does not support partial matching

2 7 2 Vector Space Model

The vector space model (Salton, 1971, Salton et al , 1975) represents documents and queries as high dimensional vectors in which each dimension corresponds to the importance of a language term The relevance of a document to a query is assumed to correlate with the similarity between the query and document vector representations This model is a general similarity framework for information retrieval that requires the choice of a weighting scheme for the query and document terms and the selection of a similarity function to compare vector representations The main benefit of this model compared to the Boolean model is that it provides a ranking of the retrieved results and does not require a full match between all query and document terms It is also a relatively simple, efficient, and effective retrieval model that has been extensively tested

The most popular term weighting schemes for the vector space model are TF-IDF models (Salton and Buckley, 1988) The TF-IDF weighting scheme models the importance of a language term as a product of a function of its within document/query frequency and a function of the term's frequency within the collection The within document/query weighting function is generally a monotonically increasing function of term frequency such as raw term frequency or the log of the term frequency The collection-based weight of a term is usually an IDF weight (Inverse Document Frequency), which in general is a monotonically decreasing function of the number of documents the term is present in The rationale behind these two weighting factors for each term is that the more times a term is present in a document, the more the term reflects its aboutness, while the more documents the term is in, the less useful the term is for discriminating between relevant and non-relevant documents

We will define an example TF-IDF weighting scheme for i-th term in the j-th document, $td_{j,i}$, and also for the k-th query, $tq_{k,i}$, as

$$td_{j,i} = log(tf_{j,i} + 1) \times IDF_i \tag{39}$$

$$tq_{k,i} = tf_{k,i} \times IDF_i \tag{40}$$

$$IDF_i = \log \frac{N+1}{n_i + 0\,5} \tag{41}$$

where

$tf_{j,i}$ = the total number of occurrences of term i in document j

$tf_{k,i}$ = the total number of occurrences of term i in query k

N = the total number of documents in the collection

n_i = the total number of documents with term i in the collection

As can be seen from this example, the query and document representations do not need to use the same weighting function In our TF-IDF weighting example the query representation uses raw term frequency, while the document representation use a log based function of term frequency The log function dampens the benefit of repeated terms under the assumption that for example a document with a term repeated 20 times is not five times more about the term than a document with the term repeated 4 times

In general the similarity function is either the dot product or Cosine similarity (dot product between normalised vectors) The Cosine similarity function can be interpreted geometrically as a measure of the angle between the query and document vectors and is defined as

$$similarity(d_j, q_k) = \frac{\sum_{i=1}^{n} (td_{j,i} \times tq_{k,i})}{\sqrt{\sum_{i=1}^{n} td_{j,i}^2 \times \sum_{i=1}^{n} tq_{k,i}^2}} \tag{42}$$

where $td_{j,i}$ = the weight of term i in document j

$tq_{k,i}$ = the weight of term i in query k

n = the total number of unique terms in the collection

The main benefit of the Cosine similarity over the dot product is that it normalises the document length removing the bias towards long documents These similarity functions are very efficient as only scores for terms that are present in both the query and document need to be calculated when searching documents

Since the query is represented simply as a vector over the complete indexing language, relevance feedback can be accommodate by updating the query term weights, such as in Rocchio (1971) where the weighted centroid of relevant and non-relevant documents are added and subtracted respectively from the initial query vector representation Similar to the language modelling approach to information retrieval, the vector space model

does not prescribe what form relevance feedback should take or have any formal notion of it built into the theoretical framework

Since the vector space model is a general similarity based retrieval framework it can accommodate the language modelling approach We can represent both queries and documents as a vector of probabilities of language terms and use the relative entropy measure as the similarity function to compare representations (Lafferty and Zhai, 2001) We can view the language modelling framework as providing a principal method for deriving the document and query vector representations as well as the selection of the similarity function Furthermore, the Jelinek-Mercer language model can be expressed as a vector space model with TF-IDF-like weighting scheme and dot product similarity function (Hiemstra, 1998, Hiemstra and Kraaij, 1999) This indicates that smoothing with the background collection in the language modelling approach to information retrieval plays a similar role as IDF in the TF-IDF weighting scheme

2 7 3 Probabilistic Models

Similar to the Boolean model, probabilistic models assume that documents are either members of the relevant or non-relevant sets for a given information need, but unlike the Boolean model, documents are ranked by their probability of being a member of the relevant set of documents The *Probability Ranking Principle* (Robertson, 1977), which is the basis of the probabilistic retrieval models, suggests ranking documents by decreasing probability of relevance, $\mathbb{P}(r|Q, D)$, and is optimal for many ad hoc information retrieval evaluation measures

In this section, we will describe the classical probabilistic model, the Robertson-Sparck Jones model (Robertson and Sparck Jones, 1976), and the more recent BM25 model, which is a very effective probabilistic retrieval model that represents the current state-of-the-art for text retrieval The probability-based approaches to information retrieval began over 40 years ago (Maron and Kuhns, 1960) and are described more extensively in (van Rijsbergen, 1979, Crestani et al , 1998) The main difference between these approaches and the language modelling approach is that the probabilistic models explicitly have the probability of relevance as their ranking criteria whereas the language models declare the generative probabilities (e g probability of the document representation generating the query sample) as their explicit ranking criteria This loss of the explicit notion of relevance in the standard language modelling approach puts it at a disadvantage compared to the probabilistic models, which can more formally accommodate relevance feedback (Sparck Jones et al , 2002)

It is further argued by Sparck Jones et al (2002) that a significant disadvantage of the language modelling approach compared to probabilistic models is that it assumes that a

single ideal/relevant document generates the query A counter argument by Lafferty and Zhai (2003) is that the query-likelihood language modelling approach is formally justified from the *Probability Ranking Principle*, and therefore both approaches to information retrieval are probabilistically equivalent This is contentious as probabilistic equivalency comes about through a weaker interpretation of relevance than originally intended by its proponents Relevance is reinterpreted as being related to a specific query instance (e g query terms) than to the underlying information need for probabilistic equivalence to be derived between models We will now present the derivations of both the Robertson-Sparck Jones Probabilistic Model and query-likelihood language model, as derived by (Lafferty and Zhai, 2003)

2 7 3 1 Robertson-Sparck Jones Probabilistic Model

In this section we present the derivation of the Robertson-Sparck Jones probabilistic model (Robertson and Sparck Jones, 1976), which ranks documents by the odds of relevance Since the odds of relevance, the ratio of the probability of relevance to the probability of non-relevance, is a monotonic transformation of the probability of relevance, this criteria is justified from the Probability Ranking Principle For query Q_k the document D_j is ranked using the following[1]

$$RSV_{k,j} = \mathbb{P}(r|D_j, Q_k) \propto \frac{\mathbb{P}(r|D_j, Q_k)}{1 - \mathbb{P}(r|D_j, Q_k)} = \frac{\mathbb{P}(r|D_j, Q_k)}{\mathbb{P}(\overline{r}|D_j, Q_k)}, \tag{43}$$

where r denotes the relevance event for the information need as expressed in the k-th query By applying Bayes' Rule, $\mathbb{P}(R|D_j, Q_k) = \mathbb{P}(D_j, Q_k|R) \times \mathbb{P}(R)/\mathbb{P}(D_j, Q_k)$, this ranking can be transformed into probabilities conditioned on the relevant and non-relevant events

$$RSV_{k,j} = \frac{\mathbb{P}(D_j, Q_k|r)\mathbb{P}(r)}{\mathbb{P}(D_j, Q_k|\overline{r})\mathbb{P}(\overline{r})} \tag{44}$$

In this section, Equation (44) is factored to produce the Roberston-Sparck Jones probabilistic model, while in the next section an alternative factoring of this equation produces the language modelling approach By factoring $\mathbb{P}(D_j, Q_k|R) = \mathbb{P}(Q_k|R)\mathbb{P}(D_j|Q_k, R)$ this becomes

$$\begin{aligned} RSV_{k,j} &= \frac{\mathbb{P}(D_j, |Q_k, r)\mathbb{P}(Q_k|r)\mathbb{P}(r)}{\mathbb{P}(D_j, |Q_k, \overline{r})\mathbb{P}(Q_k|\overline{r})\mathbb{P}(\overline{r})} \\ &= \frac{\mathbb{P}(D_j, |Q_k, r)}{\mathbb{P}(D_j, |Q_k, \overline{r})} \times \frac{\mathbb{P}(r|Q_k)}{\mathbb{P}(\overline{r}|Q_k)} \end{aligned} \tag{45}$$

Removing the query and document independent factors that do not effect the ranking of documents produces the following

$$RSV_{k,j} = \frac{\mathbb{P}(D_j|Q_k, r)}{\mathbb{P}(D_j|Q_k, \overline{r})} \tag{46}$$

[1] We follow the notation and derivation of Lafferty and Zhai (2003)

The Robertson-Sparck Jones probabilistic model ranks documents based on the ratio of the probability that the document is a sample (or generated) from the relevant set (relevance language model) compared to the probability that it is sampled (or generated) from the non-relevant set (non-relevance language model)

2 7 3 2 Relationship to language models

The language modelling approach is related to the probabilistic model as it can be derived from the same basis (Lafferty and Zhai, 2003) Note that this equivalency is achieved by a weaker interpretation of relevance (r) than in the original probabilistic model In particular, it does not support two query representations that have the same terms to have different relevance information, while the original derivation does By factoring the odds ratio in equation (44) differently, $\mathbb{P}(D_j, Q_k|R) = \mathbb{P}(D_j|R)\mathbb{P}(Q_k|D_j, R)$ it is possible to derive the query-likelihood language model [2]

$$\begin{aligned} RSV_{k,j} &= \frac{\mathbb{P}(Q_k|D_j, r)\mathbb{P}(D_j|r)\mathbb{P}(r)}{\mathbb{P}(Q_k|D_j, \overline{r})\mathbb{P}(D_j|\overline{r})\mathbb{P}(\overline{r})} \\ &= \frac{\mathbb{P}(Q_k|D_j, r)}{\mathbb{P}(Q_k|D_j, \overline{r})} \times \frac{\mathbb{P}(r|D_j)}{\mathbb{P}(\overline{r}|D_j)} \end{aligned} \tag{47}$$

If we assume that the document and query are independent when conditioned on the non-relevant event, $\mathbb{P}(D_j, Q_k|\overline{r}) = \mathbb{P}(D_j|\overline{r})\mathbb{P}(Q_k|\overline{r})$, then this leads to

$$RSV_{k,j} = \frac{\mathbb{P}(Q_k|D_j, r)}{\mathbb{P}(Q_k|\overline{r})} \times \frac{\mathbb{P}(r|D_j)}{\mathbb{P}(\overline{r}|D_j)} \tag{48}$$

For document ranking purposes we can simplify by keeping only the terms that relate to the document

$$RSV_{k,j} = \mathbb{P}(Q_k|D_j, r) \times \frac{\mathbb{P}(r|D_j)}{\mathbb{P}(\overline{r}|D_j)} \tag{49}$$

This ranking is the query-likelihood language model with priors on the odds of document relevance If we have no prior information on which to prefer different documents then this ranking is simply the query-likelihood $\mathbb{P}(Q_k|D_j, r)$ The two retrieval approaches are therefore probabilistically equivalent with this weaken form of relevance and query-likelihood could be referred to as the likelihood of the query given a relevant document but in fact the conditioning on relevance is dropped in the language modelling approach to give the standard query-likelihood function

2 7 3 3 Binary Independence Model

In the Binary Independence model (Robertson and Sparck Jones, 1976), documents and queries are represented by the set of terms they possess and it is assumed that terms are independent when conditioned on relevant and non-relevant events, or alternatively the

[2]This is the derivation from Lafferty and Zhai (2003)

assumption can be weakened slightly to linked dependencies for both events (Cooper, 1995) Therefore, the Robertson-Sparck Jones probabilistic model can be expressed in terms of the individual probability of terms conditioned on the relevant and non-relevant events, which leads to the following

$$RSV_{k,j} = \frac{\mathbb{P}(D_j|Q_k,r)}{\mathbb{P}(D_j|Q_k,\bar{r})} = \prod_i \frac{\mathbb{P}(T_{j,i}=t_{j,i}|Q_k,r)}{\mathbb{P}(T_{j,i}=t_{j,i}|Q_k,\bar{r})} \tag{50}$$

where T_j is the binary random vector for document j and $t_{j,i}$ is either 1 or 0 depending on whether the i-th term is present or absent from the j-th document We can see that this is essentially two Multiple-Bernoulli language models, one for the relevant set of documents and the other for the non-relevant set of documents Recall that Multiple-Bernoulli language models were used in the first language modelling approach to information retrieval (Ponte and Croft, 1998), which represented each document using a Multiple-Bernoulli distribution and ranked documents using query-likelihood The RSV for the Binary Independence model can be more efficiently calculated based on only the terms present in the document, by pre-multiplying under the assumption of no matching terms and then correcting this for each matched term, leading to the following relevance weight for matched terms

$$RSV_{k,j} = \prod_{t_{j,i}} \frac{\mathbb{P}(T_{j,i}=0|Q_k,r)}{\mathbb{P}(T_{j,i}=0|Q_k,\bar{r})} \prod_{t_{j,i}=1} \frac{\mathbb{P}(T_{j,i}=1|Q_k,r)}{\mathbb{P}(T_{j,i}=1|Q_k,\bar{r})} \times \frac{\mathbb{P}(T_{j,i}=0|Q_k,\bar{r})}{\mathbb{P}(T_{j,i}=0|Q_k,r)}, \tag{51}$$

$$\propto \prod_{t_{j,i}=1} \frac{\mathbb{P}(T_{j,i}=1|Q_k,r)}{\mathbb{P}(T_{j,i}=1|Q_k,\bar{r})} \times \frac{\mathbb{P}(T_{j,i}=0|Q_k,\bar{r})}{\mathbb{P}(T_{j,i}=0|Q_k,r)} \tag{52}$$

Given the set of relevant and non-relevant documents it is possible to estimate the required probabilities for the Binary Independence model using the maximum likelihood estimate We will use the following notation in this section for the k-th Query, there is R_k relevant documents in the relevant set, N documents are in the total search collection and a given term t_i is in n_i documents, $r_{i,k}$ of which are relevant for k-th query The maximum likelihood estimates are therefore

$$\mathbb{P}_{ML}(T_i=1|Q_k,r) = \frac{r_{i,k}}{R_k} \tag{53}$$

$$\mathbb{P}_{ML}(T_i=0|Q_k,r) = 1-\frac{r_{i,k}}{R_k} = \frac{R_k-r_{i,k}}{R_k} \tag{54}$$

$$\mathbb{P}_{ML}(T_i=1|Q_k,\bar{r}) = \frac{n_i-r_{i,k}}{N-R_k} \tag{55}$$

$$\mathbb{P}_{ML}(T_i=0|Q_k,\bar{r}) = 1-\frac{n_i-r_{i,k}}{N-R_k} = \frac{N-R_k-n_i+r_{i,k}}{N-R_k} \tag{56}$$

The problems with the maximum likelihood estimate of language model parameters are present in this classical probabilistic model with problems occurring if either of these probabilities are zero, which causes the odds ratio to either evaluate to 0 or infinity The solution employed for the Binary Independence model is to smooth the maximum

likelihood estimate using Lidstone smoothing with typically $\lambda = 0.5$ (Jeffrey's prior), which puts a uniform prior on a term being generated or not from the relevant and non-relevant sets

$$\mathbb{P}_{Lid}(T_i = 1|Q_k, r, \lambda) = \frac{r_{i,k} + \lambda}{R_k + 2 \times \lambda} \quad (57)$$

$$\mathbb{P}_{Lid}(T_i = 0|Q_k, r, \lambda) = \frac{R_k - r_{i,k} + \lambda}{R_k + 2 \times \lambda} \quad (58)$$

$$\mathbb{P}_{Lid}(T_i = 1|Q_k, \overline{r}, \lambda) = \frac{n_i - r_{i,k} + \lambda}{N - R_k + 2 \times \lambda} \quad (59)$$

$$\mathbb{P}_{Lid}(T_i = 0|Q_k, \overline{r}, \lambda) = \frac{N - R_k - n_i + r_{i,k} + \lambda}{N - R_k + 2 \times \lambda} \quad (60)$$

By plugging these estimates into the Equation 52, we can rank the documents using

$$RSV_{k,j} = \frac{\mathbb{P}(D_j|Q_k, r)}{\mathbb{P}(D_j|Q_k, \overline{r})} \propto \prod_{t_j, i=1} \frac{r_{i,k} + \lambda}{n_i - r_{i,k} + \lambda} \times \frac{N - R_k - n_i + r_{i,k} + \lambda}{R_k - r_{i,k} + \lambda} \quad (61)$$

Obviously in practice relevance information is not available for the full collection of documents and therefore the relevance feedback in this model is based on a subset of relevance ratings of documents that is either supplied by a user or by the assumption that a specific number of the initial top ranking results are relevant

In this retrieval model, when we do not know any relevant documents we set $P(t_i|Q_k, r)$ to a constant and $P(t_i|Q_k, \overline{r})$ is approximated with the probabilistic IDF, which leads to the following heuristic approximations

$$P(t_i|Q_k, r) = 0.5 \quad (62)$$

$$\mathbb{P}(t_i|Q_k, \overline{r}) = \log \frac{N - n_i + 0.5}{n_i + 0.5} \quad (63)$$

The benefit of this model over language models is its inherent support for relevance feedback, while its weakness is its support for initially ranking documents without relevance information, which is done heuristically in this model and is better supported by language models due to their better statistical estimation techniques coupled with the query-likelihood function Also, as discussed previously the Multiple-Bernoulli probability model represents documents and queries as binary vectors and therefore only provides limited representation compared to the more common multinomial probability distribution that is used in the language modelling approach to information retrieval

2.7.3.4 Okapi BM25 Retrieval Model

The Okapi BM25 retrieval model (Robertson et al., 1995), which approximates the 2-Poisson model (Bookstein and Swanson, 1974, Harter, 1975), is a more effective probabilistic model than the Binary Independence model The BM25 model has a richer

representation of documents as a stream of terms in comparison to the Binary Independence model, which simply represents documents as a binary vector of terms

The basis of the BM25 model, the 2-Poisson model, assumes that terms occur within documents as a mixture of two Poisson distributions, which represent the probability of elite and non-elite occurrences of terms When terms occur due to eliteness, it is assumed that the document is about the term, otherwise it is a non-elite term occurrence and the document is not assumed to be about the term's concept While the BM25 model has its theoretical foundations in the 2-Poisson model, its actual realisation is largely guided by good IR practice and experimentation and the BM25 model cleverly approximates the 2-Poisson model and incorporates both the Robertson-Spark Jones weight (Robertson and Sparck Jones, 1976) and document length normalisation

The BM25 retrieval model is defined as

$$BM25 = \sum_{t_i \in Q} \log \frac{(r_i + 0\,5)(N - R - n_i + r_i + 0\,5)}{(n_i - r_i + 0\,5)(R - r_i + 0\,5)} \times \frac{(k_1 + 1)tf}{K + tf_i} \times \frac{(k_3 + 1)qtf_i}{k_3 \times qtf_i} \quad (64)$$

where $K = k_1((1 - b) + b \times dl/avdl))$, tf_i is the term frequency within the document, qtf_i is the term frequency within the query, dl is the document length and $avdl$ is the average document length The parameters k_1, k_3, b depend on the nature of the queries and the collection and can be optimised for example to suit small or large queries The k_1 parameter controls the rate of increase of a document's term weight in response to increasing frequency and the k_3 parameter likewise controls the query term weight, while the b parameter controls the effect of document length normalisation on ranking

The BM25 model is extensively tested on the TREC test collections and has previously been utilised for video shot retrieval in the Informedia's TRECVid experiments (Hauptmann et al , 2004) In general this retrieval model performs very well on the TREC test collections For this reason, in this thesis we will compare different language modelling approaches with the BM25 model as opposed to the Binary Independence model for the text-based video shot retrieval task

2.8 Summary

Language models are simply generative probability distributions for text sequences that can be used in information retrieval to represent both documents and queries The standard language modelling approach to information retrieval models documents with unigram multinomial distributions that are smoothed using the background collection model and ranks documents based on query-likelihood, the probability of the document's language model generating the query

Smoothing is a common method for improving the estimation of language models

especially for missing and low frequency words and can be accomplished by either discounting, combination, back-off or validation methods Discounting methods smooth unobserved words uniformly, while the more popular combination-based methods combine the maximum likelihood estimate of the document's text with a background model and are the most effective smoothing methods for the text retrieval task Back-off strategies work consistently poorer than combination-based methods and validation methods can be used to automatically set the parameters of other smoothing methods The most popular smoothing method for text retrieval is Jelinek-Mercer smoothing which combines the document's maximum likelihood model with the collection model in a simple linear interpolation (finite mixture model) The Dirichlet language model is also very effective and combines the background collection model as a prior distribution with the document model in an empirical Bayes framework

The language modelling approach to information retrieval generally either ranks documents by query-likelihood or by relative entropy Query-likelihood is the most common ranking approach in LMIR and ranks documents by their probability of generating the query Whereas more recent approaches use relative entropy (or equivalently cross entropy), which ranks documents by the difference in information between the query language model and the document language model The advantage of relative entropy is that it more naturally supports relevance feedback and also has the nice property that if the query model is the maximum likelihood estimate then it will produce exactly the same ranking as the query-likelihood method The original query-likelihood ranking method severely constrained LMIR approaches in their support for relevance feedback whereas relative entropy supports a far more powerful query representation that is taken advantage of in the current approaches to relevance feedback

The relative entropy ranking approach in LMIR has similarities with the vector space model but mimics many of its features in a principled way, for example smoothing can be seen to be related to IDF in reducing the effect of common words The language modelling approach also has similarities with the probabilistic IR approaches as it can be derived from the same principle, the Probability Ranking Principle, that forms the basis of the probabilistic models However, the lack of explicit relevance in the generative language modelling approach makes it theoretically less principled for supporting relevance feedback The main contribution of the LMIR approach is the intricate statistical estimation techniques and the simple generative framework, which can as easily be applied to information retrieval in other media such as image, video and audio as it is applied to text retrieval

Language modelling-based information retrieval is an active area of research within the information retrieval community and since its introduction to the IR field in 1998, a series of papers has been published yearly at most major IR conferences, numerous LMIR workshops have been organised and even a book on the topic has been published

(Croft and Lafferty, 2003) Beyond ad hoc text retrieval LMIR is also being applied to video retrieval (Westerveld, de Vries and van Ballegooij, 2003), topic tracking (Kraaij and Spitters, 2003) and cross-language retrieval (Xu et al , 2001)

CHAPTER III

STATE-OF-THE-ART IN VIDEO RETRIEVAL

Video retrieval combines methods from text retrieval, concept-based retrieval, visual-based retrieval and to a lesser extent audio-based retrieval for the task of locating video clips that meet a user's information need from video collections The overall performance of current video retrieval systems heavily depends on text retrieval but low-level visual retrieval based on colour, edge and texture, while less semantically rich than text, provides support for query-by-example visual searching Research into retrieval based on high-level audiovisual concepts seeks to bridge the gap between low-level audiovisual features and the high-level information needs of real users Due to the multimodal nature of video documents and queries, fusion of retrieval results from different feature representations plays a significant role in determining the effectiveness of video retrieval Relevance feedback can also provide a mechanism to improve the video query representation and the fusion of features Benchmarking of retrieval techniques and features is integral in progressing video retrieval research and is supported by the recent TRECVid initiative

3 1 Introduction

3 2 Video Indexing

- 3 2 1 Structuring
- 3 2 2 Text
- 3 2 3 Concepts
- 3 2 4 Visual
- 3 2 5 Audio

3 3 Video Retrieval

- 3 3 1 Query Preprocessing
- 3 3 2 Text-Based
- 3 3 3 Concept-Based
- 3 3 4 Visual-Based
- 3 3 5 Fusion
- 3 3 6 Relevance Feedback

3 4 Evaluation

3 5 Summary

3.1 Introduction

Video retrieval research is concerned with improving users' access to video collections through the development of better retrieval systems and interfaces The objective is to improve the video retrieval experience for real users interacting with real video collections in terms of system effectiveness and efficiency for the tasks they wish to accomplish and of course in terms of user satisfaction In this thesis we focus on general ad hoc retrieval by a hypothetical professional user, such as one concerned with finding video footage for news or documentary programmes Furthermore, since this thesis concerns retrieval models, we will discuss retrieval and fusion models in this chapter in preference to user models, user interfaces, user-satisfaction or other human computer interaction factors

Content-based video retrieval is a recent application that has been facilitated by improvements in computing power networks, storage and video compression techniques The visual techniques in practice in video and image retrieval often originate from research in Computer Vision applications (Rui, Huang and Chang, 1997), while the retrieval models and particularly more recently the evaluation methodology is heavily influenced by the information retrieval field In the early 90s research on content-based image retrieval began in earnest to facilitate the search of image collections by their content without the need for costly manual annotation Early image retrieval systems include QBIC (Flickner et al , 1995), Virage (Bach et al , 1996), Photobook (Pentland et al , 1996), VisualSEEk/WebSEEk (Smith and Chang, 1996b, 1997), Netra (Ma and Manjunath, 1997) and PicHunter (Cox et al , 2000) Content-based video retrieval research followed with extensions to some of these image retrieval systems and also with the development of video retrieval systems such as MARS (Ortega et al , 1997), VideoQ (Chang et al , 1997), Video Mail Retrieval (Brown et al , 1995) and Informedia (Wactlar, 2000)

This early phase of video retrieval research has been enriched by the addition of common video collections and search tasks organised initially by NIST's TREC-10 Video Track in 2001 This first TREC video track was later followed by annual TRECVid workshops that distributed larger and higher quality video collections In this chapter we will study the approaches to video retrieval that are for the most part tested on the TRECVid collections and which are currently focussed on TV news and documentary style material Previous approaches to evaluation of image retrieval systems are limited in applicability to this type of retrieval scenario This is in part due to the small size of collections involved in those evaluations but more significantly because of differences in the type of retrieval task being evaluated which often concerned texture retrieval or classification of images into homogeneous groups

The best single feature for effective retrieval of video is text, particularly text from automatic speech recognition (ASR) or closed caption text, at least for TRECVid collections consisting of broadcast TV news programmes It is worth noting that the dominance of text would be even greater if the TRECVid topics were less skewed towards visual retrieval When video retrieval is viewed as a text retrieval problem it seems rather small scale - video retrieval systems (and test sets) are many orders of magnitude smaller than the TREC Terabyte collection (Clarke et al , 2005) in terms of their text content The recent TRECVid search test collections consist individually of nearly 3MB of ASR text (33,000 shot documents) compared to the TREC Terabyte search collection which contains 426GB of HTML text (25 million documents) However, video is more than just text and the challenge is to support users performing visually-oriented retrieval and to even improve on text-based retrieval by fusing it with results from audiovisual retrieval models

Content-based video retrieval systems require a significant indexing effort in order to produce content descriptions that support retrieval This indexing phase involves structuring the video content and extracting audiovisual features such as text, colour, edge, texture, motion, audio features and others The recent MPEG7 standard (Multimedia Content Description Interface) standardises these content descriptions to improve interoperability between video retrieval systems The standard is somewhat of a guideline for efficient video retrieval using compact video features The choice of audio and visual features in the MPEG7 standard has been guided by the major research groups involved in this research area and takes on board many years of experience and experiments in the video retrieval field but it should be noted that many of the image and video experiments are based on small collections or tasks that bear little resemblance to general ad hoc video retrieval Many but not all of the successful visual features are based on the histogram (or a histogram-like feature), which is essentially a discrete probability distribution and this fact will allow us in future chapters to frame content-based retrieval in terms of the (discrete) language modelling approach to information retrieval Current retrieval systems typically use geometric distances such as L1 or L2 for similarity search on histogram features With such a wide variety of features available within content-based video retrieval systems, fusion plays a significant role in the overall effectiveness of any video retrieval system

The rest of the chapter is organised as follows In Section 3 2 we describe video indexing which consists of structuring the video and extracting features to represent the video content We follow this in Section 3 3 with descriptions of the current state-of-the-art in video retrieval which consists of text-based retrieval, concept-based retrieval, visual-based retrieval, fusion and relevance feedback In Section 3 4 we describe the TRECVid evaluation framework for ad hoc video search Finally, in Section 3 5 we summarise our study of the current state-of-the-art in video retrieval

3.2 Video Indexing

Video indexing is the process of preparing video collections for retrieval Operational video retrieval systems use different levels of human involvement in the indexing process such as for categorisation of the content, adding rights management and structured descriptions of constituent video segments In this thesis we concern ourselves with fully automatic indexing techniques Fully automatic indexing tools complement these manual and semi-manual efforts by providing alternative retrieval functionality (e g visual query-by-example) as well as lessening the workload on manual indexers if integrated into semi-automatic indexing tools

Automatic video indexing involves supplementing any existing descriptions of the

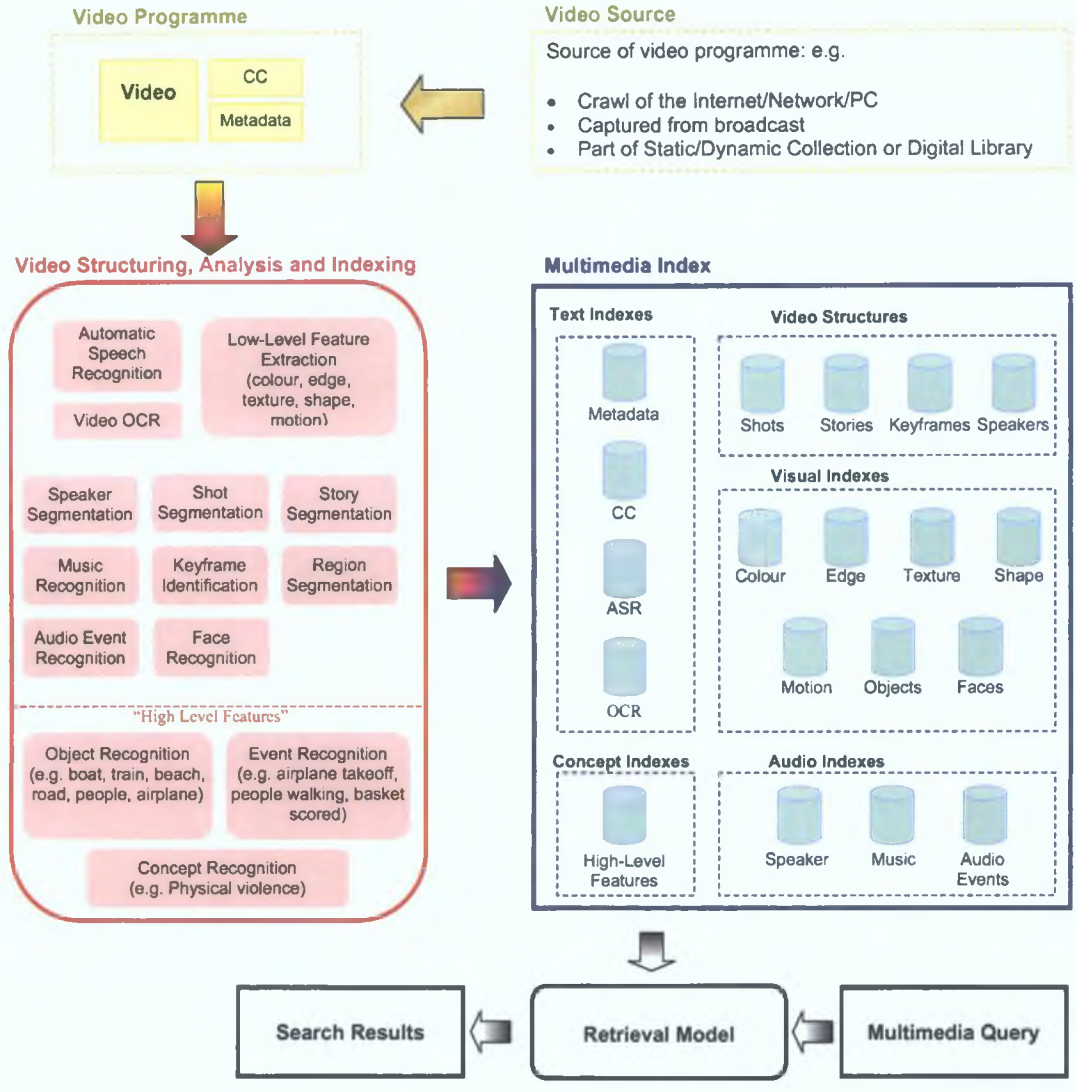


Figure 4: Overview of video indexing system.

content through content structuring and feature extraction, and creating efficient indexes to support browsing and query-based retrieval. In our discussion of this area we will primarily discuss content structuring and feature extraction methods, while indexing structures such as efficient multidimensional indexes are outside the scope of our research. The purpose of video structuring is to expose the video's semantic and physical structure (e.g. shot and story segmentation) so as to make it easier to index, browse and search, while feature extraction creates descriptions of the video content that are effective for retrieval. Zhang et al. (1995) present early research on video indexing (video parsing in their parlance) which they define as involving two key processes – "the *temporal segmentation* of a video programme into its elemental units, and *content extraction* from those units, based on both video and audio semantic primitives."

A typical video indexing system is outlined in Figure 4. Videos that are indexed originate from a source that invariably supplies some initial information. If the video

collection is from a crawl of the internet then filenames, text from related web pages that link to the content, and timestamps provide an initial description of the content If the video is captured from broadcast television then the video can be supplemented with closed caption (CC) text encoded in the broadcast for the hearing impaired and programme information such as title, channel, timing and descriptions from the electronic programme guide for the captured programme Finally, if the source is a static collection, then existing descriptions may be available for the content in a variety of proprietary or open standards such as MPEG7 A television archive may also have manual annotations and production information such as edit decision lists which identify the shot cuts available for its video content

The video indexing process converts the initial video description of the content into a fuller description that better supports retrieval The video is broken up into retrieval units such as scenes, shots, speaker segments that structure the content into more useful units for retrieval and indexing Image and video segments (keyframes and key-segments) can be further extracted to represent these retrieval units Features such as the speech text, on screen video text and low-level visual features such as colour, edge, texture, motion and shape are extracted from these representations of the video in order to create useful content-based indexes to search High-level features such as faces, people, animals and other specific types of objects and events can also be extracted with varying degrees of accuracy High-level audio features can be defined such as monologue, dialogues and music Feature extraction and content analysis tools provide the information that forms the basis of the indexes in video retrieval (see Figure 4) The output of the video structuring and extraction indexing process can be expressed in MPEG7 (MPEG7 Committee, 2002), which is a recent international ISO standard for audiovisual content description

The MPEG7 standard, Multimedia Content Description Interface, defines the syntax and semantics of video descriptions Previous MPEG (Moving Pictures Expert Group) standards such as MPEG1 (video on storage media such as VCD), MPEG2 (video for digital television and DVD) and MPEG4 (even higher compression of digital video that supports object-based encoding) concerned the encoding of the audiovisual signal, whereas the MPEG7 standard concerns the description of the video The MPEG7 standard allows descriptions of video to be interoperable between video retrieval systems It also provides a clean interface to individual video indexing tools which can be viewed as a functional black box that takes as input the video, its initial MPEG7 descriptions and outputs the updated MPEG7 description

The MPEG7 standard is broken up into five normative parts

- Part 1 MPEG7 Systems - binary format for encoding MPEG7 descriptions and terminal architecture

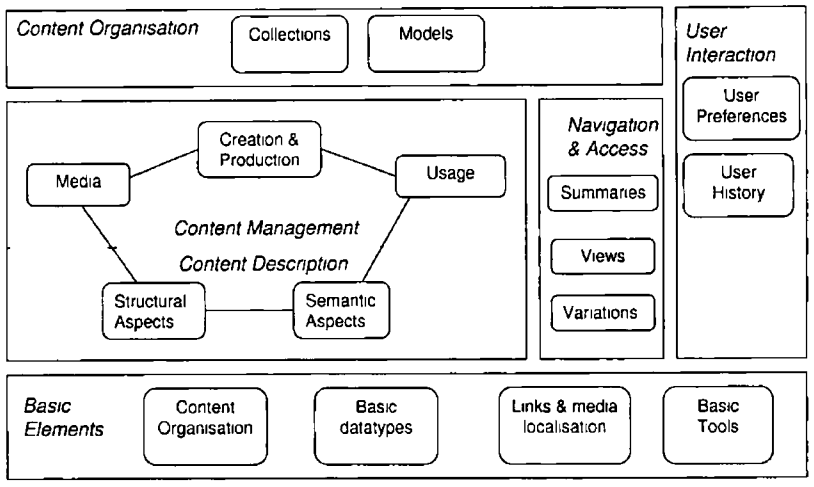


Figure 5 Overview of the MPEG7 Multimedia description schemes, originally published in (MPEG7 Committee, 2002)

- Part 2 MPEG7 Description Definition Language (DDL) - extensions to XML-Schema for audiovisual descriptions
- Part 3 MPEG7 Visual - descriptions for visual features such as colour, texture, shape, motion, localisation and face recognition
- Part 4 MPEG7 Audio - descriptions for audio features
- Part 5 MPEG7 Multimedia Description Schemes (MDS) - general descriptions for content, its management, organisation, navigation, access and also user interaction (see Figure 5)

Descriptions within the MPEG7 standard are either Descriptors or Description Schemes Descriptors represent features within the MPEG7 standard, while Description Schemes combine descriptors and other Description Schemes The MPEG7 Multimedia Description Schemes (MDS) defines the overall structure of an MPEG7 description for audiovisual content The main components of the MDS are shown in Figure 5 The MDS allows content to be decomposed both temporally and spatially, thereby allowing description of sub-units such as shots, objects or regions In Figure 6 we illustrate temporal, spatial and spatio-temporal segmentations supported by the MPEG7 standard Physical sub-units of the content may be described using the MPEG7 Visual and MPEG7 Audio parts of the standard, which define the Descriptors and Description Schemes for low-level visual and audio features The MPEG7 description format can be extended using the MPEG7 Description Definition Language (essentially XML-Schema) but extensions to the MPEG7 descriptions should be discouraged due to the inevitable incompatibility it introduces between consumers of the descriptions The MPEG7 System tools provide a mechanism for the MPEG7 standard, which is XML based, to be encoded in a compact binary representation and supports multiplexing and synchronising the description with the video content

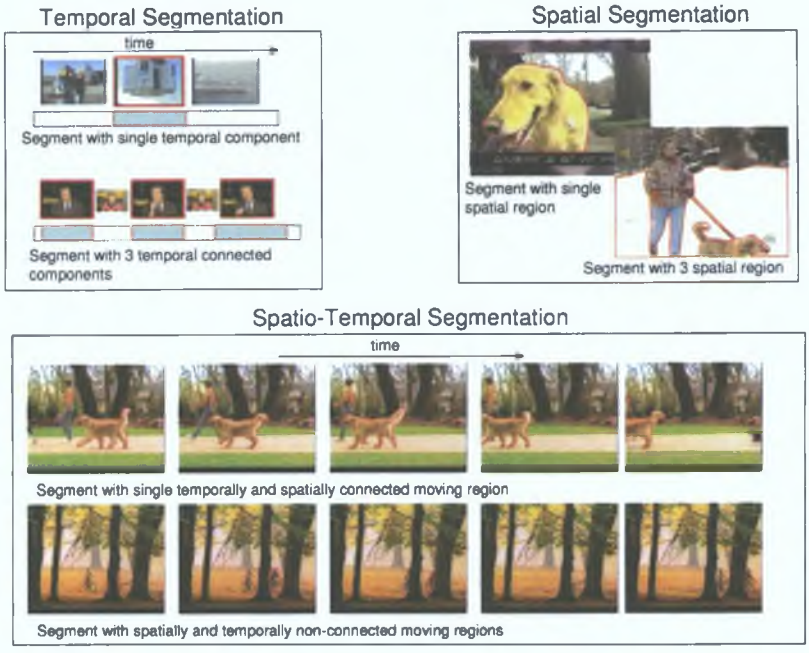


Figure 6: Segmentation of the audiovisual content.

In the following sections we describe the process of structuring video and extracting features such as text, high-level concepts, visual features and audio features. We highlight the support within the MPEG7 standard for each of these indexing processes.

3.2.1 Structuring Video

Content-based retrieval requires that the video content be structured into meaningful and useful sub-units so as to better support content-based browsing and querying. There are many video segmentations possible for video content as illustrated in Figure 7. The video can be decomposed into stories, shots and speaker segments. Story segmentation is particularly useful for news videos since each story is usually independent of other stories in the same programme. Both shot and speaker segmentations are physical segmentations that can be achieved with higher accuracy than story detection. A shot, a continuous sequence from a camera, is a useful retrieval unit since it is visually coherent. A speaker segment, an uninterrupted sequence of utterances from a single speaker, provides a consistent audio unit. The shot and speaker segmentations are equally applicable to non-news content but stories, semantically separated segments, are not as strongly defined in other types of content. The semantic unit for non-news content may be a group of shots from the same scene or a group of shots from a DVD chapter when the source content is from a DVD.

Instead of text-based retrieval models being based on shots some video systems utilise speaker segmentation as the structure for text retrieval (Chua et al., 2005) or as part of a text retrieval structure (Rautiainen et al., 2005). Shots provide useful support for

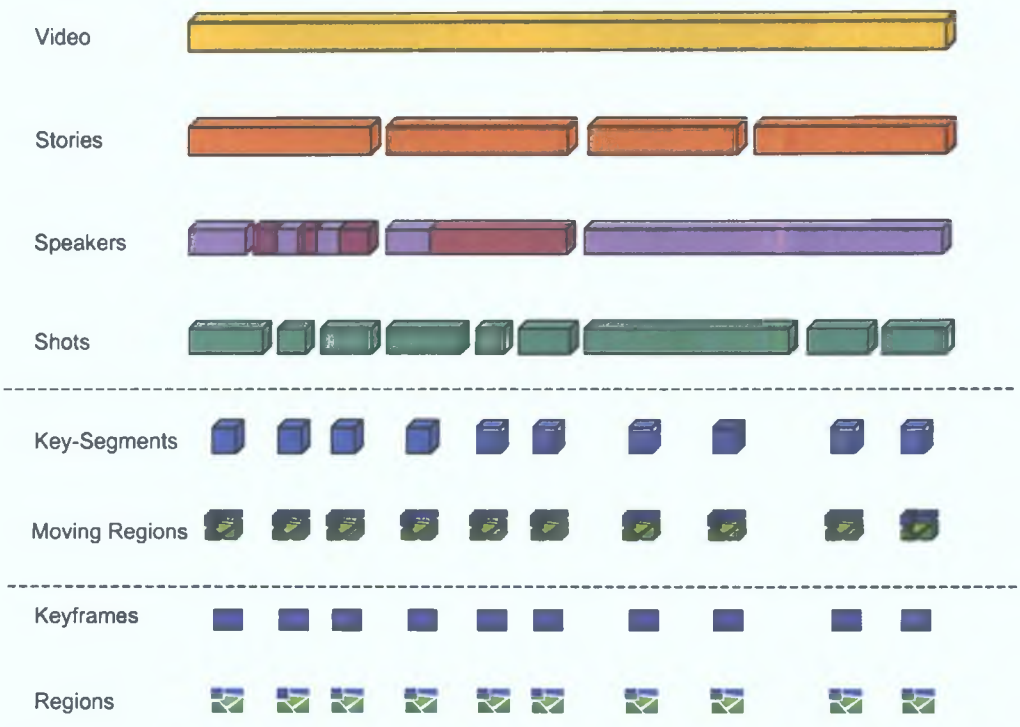


Figure 7: Illustration of the different segmentations of a television news programme.

Figure 8: Examples of different shot transitions.

queries seeking video that shows some item of interest. General information seeking in news video would of course be better supported by news story segmentation, which identifies each story within a news programme, as it provides a better browsing structure and grouping structure for seeking information.

A more detailed segmentation of the video is required to support visual retrieval. Shots can be further decomposed into key-segments, keyframes and regions. Features are then extracted from these units that act as the index of the retrieval unit when performing similarity or sketch-based querying. A key-segment is a temporal segment, while a keyframe is a single image from a shot. Most current video retrieval systems represent a shot by simply selecting a single keyframe. Key-Segments can be further decomposed into spatio-temporal regions that are visually coherent and likely to represent some object or part of one in space and time. Likewise, keyframes can be further segmented into regions of the image that are visually coherent. This provides an even finer level of retrieval for objects or parts of objects that whole keyframe or key-segment representations cannot support. Unfortunately, region segmentation of general video content does not segment video frames into semantically meaningful regions and therefore is currently of limited usefulness in video retrieval.

Only a few video retrieval systems extract spatio-temporal features beyond motion. Rautiainen et al. (2005) create their spatio-temporal representation of video shots by sampling 20 video frames evenly over the bounded video sequence, and extract two visual features, a Temporal Colour Correlogram (Rautiainen and Doermann, 2002) and a Temporal Gradient Correlogram (Rautiainen, Seppanen, Penttila and Peltola, 2003). Westerveld et al. (2004) extract a one second spatio-temporal segment centred on the keyframe and model these 29 frames after converting each frame using the DCT transform as a Gaussian Mixture Model. In nearly all other TRECVid video shot retrieval systems visual features (except motion) are extracted on keyframes.

3.2.2 Text Descriptions

There are three possible text indexes available within a video retrieval system – the automatic speech recognition text (ASR text), the video optical character recognition text (video OCR text) and the closed caption text (CC text). The ASR text is a transcript of what is spoken. The video OCR text is a transcript of text that is visible within the video, which is commonly used in interviews and news reports to identify people, their title and location. Video OCR text can also contain text recognised from background objects within the frame and from advertisements. The CC text is a representation of what is spoken that is transmitted with the television programme as an aid for the hearing impaired. It is not normally a word-by-word transcription of what is spoken and sometimes includes change of speaker colour codings and identification of some audio

events for general programmes (e g knock at door, phone rings)

These three source of text can complement each other The ASR is the fullest account of what is spoken within the video The CC text can contain the proper spelling of named entities that may be incorrectly recognised in the ASR text due to being out of vocabulary terms The video OCR text is also useful as the text is often directly related to the people or location that is present in the video shot, whereas items mentioned in ASR text may not be present in the related shot

The standard IR text preprocessing is usually applied to the ASR, video OCR and CC text such as stopword removal and stemming These texts are easily aligned to shots as they are timestamped relative to the video In the case where the CC timestamps are missing, then they may be aligned with the video based on the ASR transcript (Rautiainen et al , 2005, Ratcliff and Metzener, 1988) Even though the text resources are synchronised with the video, there is of course no guarantee that the items mentioned in the texts are visible in the related shots

Implicit in the use of text for video retrieval is that the spoken text and visual text indicates the visual content However, due to a time-delay problem, where concepts mentioned in the audio are not immediately visible in the corresponding shot, naively representing shots by only their ASR text can be quite ineffective

Cheng and Chen (2005) try to correct the alignment problem for ASR nouns during content indexing by attempting to find the relationship between ASR words and the shots' automatically identified high-level concepts so as to associate the right ASR words to the right shot The nouns are identified by using a part-of-speech tagger and the distance between an ASR word and a high-level concept is the sum of the distance from a common ancestor in WordNet to each concept The majority of TRECVid video retrieval systems handle the alignment problem by some form of score propagation from adjacent shots at retrieval time

Text can easily be associated with different structural units in an MPEG7 video description by using the MPEG7 TextAnnotation description It is usual to represent ASR, CC, or video OCR text as simply an MPEG7 FreeTextDescription within the TextAnnotation description The type of text and the confidence of its correctness can also be specified

3 2 3 High-Level Concepts

Another approach to video retrieval is to build an ontology of high-level concepts (semantic features) that are useful for video search These high-level concepts can be detected offline during indexing using training data and other possibly concept-specific analysis

Table 1 High-level features evaluated in TRECVid 2002, 2003 and 2004 Note some features are repeated in successive years

TRECVid 2002	1 Outdoors, 2 Indoors, 3 Face, 4 People, 5 Cityscape, 6 Landscape, 7 Text Overlay, 8 Speech, 9 Instrumental Sound, and 10 Monologue
TRECVid 2003	11 Outdoors, 12 News subject face, 13 people, 14 building, 15 road, 16 vegetation, 17 animal, 18 female speech, 19 car/truck/bus, 20 aircraft, 21 news subject monologue, 22 non-studio setting, 23 sporting event, 24 weather news, 25 zoom in, 26 physical violence, and 27 Madeleine Albright
TRECVid 2004	28 Boat/ship, 29 Madeleine Albright, 30 Bill Clinton, 31 Train, 32 Beach, 33 Basket scored, 34 Airplane takeoff, 35 People walking/running, 36 Physical violence, 37 Road

tools making for very efficient and effective retrieval when the concept coincides with a user's video query Examples of concepts for video search are outdoors, indoors, faces, people and cityscape The high-level features that are tested in TRECVid are shown in Table 1 and for the most part these TRECVid high-level features are high-level concepts except the features concerning specific named people High-level concepts try to bridge the semantic gap between the users' video request and the low-level content-based representations

We will take a black-box approach to these high-level features and assume each feature has a list of shots with associated confidence values that reflect the belief of the concept-detector of the presence of the feature within the shot We will not concern ourselves with the details of feature detection (see Naphade and Smith (2004) for an overview of high-level feature detection approaches) High-level feature detection differs from video retrieval due to the relatively large amount of training data available and is more naturally a classification problem Since the high-level features are detected during indexing it is possible to have specific algorithms and heuristics chosen for each high-level feature General approaches to their detection use statistical models such as Support Vector Machines, Gaussian Mixture Models and Hidden Markov Models (Adams et al , 2003)

Simple binary high-level features as used in TRECVid can be assigned to video clips by using the MPEG7 KeywordAnnotation description within the MPEG7 TextAnnotation description of a clip Confidence values can be assigned to the TextAnnotation to indicate the strength of the classifier's belief that it is correctly assigned An MPEG7 Classification scheme should be created to provide a controlled list of concepts and to describe their relationships to each other

(a) Global (b) 3x3 Grid (c) 4 Corners + Centre

Figure 9: Spatial localisation of visual features. (a) No localision, (b) grid-based localisation, and (c) overlapping regions such as centre and corners.

3.2.4 Visual Descriptions

Low-level visual features such as colour, edge, texture and motion form the basis of query-by-example approaches to video retrieval. Colour is the most widely used feature and is also the most effective of the low-level features for TRECVid video shot retrieval. The texture features characterise the different spatial patterns within the video and edge features may indirectly characterise the shape of the objects within the video. High-level object-based features and even direct shape features have limited effectiveness for current video retrieval approaches since object segmentation is not yet mature enough to automatically segment real-world objects in general video collections.

The low-level visual features colour, texture and edge can be globally defined for a single keyframe, locally (spatially) defined for regular sub-regions of the keyframe, or even spatio-temporally (spatially and temporally) defined for a sequence of video frames. Global representations are the simplest and most common use of low-level features and represent the visual content of the full keyframe without localising the descriptions to specific spatial regions of the image.

Local representations of visual features can spatially localise the feature by breaking the image into regular regions (see Figure 9). The most common mechanism is to use a grid to break the image into rectangular regions and this is usually performed for either 3x3, 4x4 or 5x5 regions. It is also possible to achieve some degree of spatial localisation through overlapping rectangular regions such as the four corners of the video frame and centre of a video frame (Amir et al., 2004). Usually both the query image and video keyframes have the same grid applied and matching is performed for the whole image, essentially adding the normalised x and y indexes of the grid (or region index in the case of overlapping regions) as extra dimensions to the low-level feature representations. Alternatively, each regular region of the keyframe can be treated as a sub-image and shots can be ranked based on the best sub-image that matches the query image (Wu et al., 2004).

Image features can be extended in the temporal dimension. From a strictly low-level visual feature point of view, it is the temporal dimension that distinguishes video

retrieval from image retrieval Temporal extensions to colour, texture and edge features are calculated based on a sequence of images from a shot instead of a single keyframe In (Rautiainen et al , 2005) their spatio-temporal colour and edge feature is created by sampling 20 video frames evenly over the video shot, whereas in (Westerveld et al , 2004) their spatio-temporal texture feature is created for a one second sequence of video frames centred on the keyframe For the majority of video retrieval approaches the temporal domain plays little role in direct video retrieval

Visual features are often defined in terms of a distribution across the image or video segment, which is quite naturally represented as histograms More generally features including histograms can be physically represented as a vector describing the visual content and similarity between query and documents can be defined in terms of the distance between vector representations In fact multiple features can be combined into a single vector representation for the visual segment, which is often referred to as early fusion of visual features, and the vector may be further preprocessed before calculating distances, such as normalising its entries by their variance (Hauptmann et al , 2003) An alternative representation for features that have a continuous distribution is the Gaussian Mixture Model, which matches queries and video segments using probabilistic measures such as query-likelihood and document-likelihood (Westerveld and de Vries, 2004)

The benefit of the histogram representation of a distribution is that it is compact and efficient when the dimension of the feature being indexed is low It quickly loses this benefit for medium to high numbers of dimensions since its size increases exponentially with the number of dimensions In a multidimensional histogram the range of each dimension is divided into a fixed number of bins with fixed, though not necessarily uniform widths, and each feature point from the features distribution is assigned to the bin whose range contains it Histograms can form the basis of a discrete probability distribution though they are normally used in similarity based retrieval models with distance functions such as Manhattan and Euclidean distance

Gaussian mixture models (GMMs) are continuous probability models and represent multidimensional visual features by using a linear interpolation of Gaussian distributions that each have their own mean vector and covariance matrix The Gaussian Mixture Model is usually estimated during indexing time using the EM algorithm (Dempster et al , 1977) The advantage of GMMs is that they are more compact than histograms for medium to high dimensional data and also their parameters are learned from the sample data and therefore use their representation more wisely on patterns that are present in the sample distribution The main problem with GMMs is that they are slower to index and retrieve from than histogram approaches

In the following sections we describe colour, edge, texture, motion, face and spatio-temporal visual features that are commonly used in visual retrieval systems We also describe the MPEG7 descriptors that support each of these feature types We can view many of these features as describing visual languages of discrete symbols

3 2 4 1 Colour Descriptions

Colour is the most popular low-level visual feature in video retrieval Colour is our perception of light of different wavelengths that are within the visible spectrum (approximately 380nm - 750nm) Objects reflect and absorb different wavelengths of light and thereby we perceive them as coloured differently A light source (sun, incandescent bulb, or reflected light from say the moon) also emits light of different wavelengths and therefore the perceived colour of the same object in different lighting conditions will vary

Colour can be represented using different colourspaces such as RGB, YCbCr, HSV, XYZ, LUV, LAB, or Munsell The RGB colourspace is the hardware oriented model and is often used for rendering to the computer screen The YCbCr colourspace is another physical colourspace but used for encoded videos The HSV colourspace is a more perception-based model that separates the colour dimensions into Hue, Saturation and Value (Brightness) The XYZ colourspace is an international standard colourspace that can define any colour humans see with positive values of its three primaries The LUV and LAB colourspace are designed so that the same Euclidean distance between colours measures similarly different colours based on human perception Finally, the Munsell colourspace is more artistic-oriented and is useful for selecting and organising colours

In the MPEG7 standard colour can be expressed using either the RGB, YCrCb, HSV, HMMD or Monochrome colourspaces In some of the MPEG7 visual descriptions a particular colourspace is mandatory The Monochrome colourspace is simply the Y component of the YCbCr colourspace The MPEG7 HMMD colourspace is a nonlinear, reversible transformation from the RGB colourspace and defines five dimensions - Hue (as in HSV), Max (max of R,G,B triplet), Min (min of R,G,B triplet), Diff ($Max - Min$) and Sum ($\frac{Max+Min}{2}$) In addition a linear transformation can be specified in the MPEG7 standard that maps a colour descriptor's colourspace from the RGB colourspace The quantisation levels of the different colourspaces can also be specified

Video and image retrieval systems often use the HSV colourspace for histogram representations though many of the other colourspaces are equally valid choices Kirminen and Gabbouj (2000) compare RGB, HSV, LAB and XYZ colourspaces for image retrieval using human judgements on a small image collection (235 JPEG images) with 8 query

images For the colour histogram feature they find that HSV and RGB are superior to the LAB and XYZ colourspace and for the dominant colour feature they find that HSV and RGB are better than LAB Wan and Kuo (1996) find that HSV is superior to RGB for image retrieval

Like all visual features, colour features can be extracted globally for the full video frame, locally for spatial segments of a single video frame or for a series of temporally related video frames Localised features can be created for all colour features by simply breaking the image into a grid of sub-images The MPEG7 standard visual descriptions support very compact representation of the visual content, whereas regional grid-based colour histograms as used later in this thesis are generally many times larger than these representations

Common colour features in image and video retrieval are colour moments, dominant colours, colour sets, histograms, coherence vectors and correlograms Other global colour representations are Illumination Invariant Colour Descriptors (O'Callaghan and Bull, 2002), which were used in (Pickering et al , 2003) for video retrieval The five colour features defined in the MPEG7 standard are Dominant Colour, Scalable Colour, Colour Layout, Colour-Structure and GoF/GoP Colour (Manjunath and Ohm, 2001) Many of these colour descriptions (MPEG7 Colour Layout, MPEG7 Colour Structure, colour correlograms, coherence vectors) model in some respect the spatial distribution of the colour distribution within an image and therefore can also be considered as texture descriptions

We will now describe each of these colour features

- **Colour Moments** Colour moments provide a simple and compact representation of the colour of an image The k-th moment M_i^k for the i-th band of the colourspace is defined as

$$M_i^1 = \frac{1}{N}\sum_{j=1}^{N} p_{ij} \text{ , if } k = 1, \tag{65}$$

$$M_i^k = \left(\frac{1}{N}\sum_{j=1}^{N}(p_{ij} - M_i^1)^k\right)^{\frac{1}{k}} \text{ , otherwise,} \tag{66}$$

  where p_{ij} is one of N pixels in the image The first moment (average colour) and sometimes the second (standard deviation) and the third (skewness) are used in image retrieval These descriptions are ineffective for describing the colour of a whole image for video retrieval but are more appropriate for describing the colour of small or homogenous image regions The first moment (average colour) was extracted using the Munsell colourspace to describe whole images and segmented objects in the QBIC system (Flickner et al , 1995) The first colour moment forms a bound on the more expensive histogram computation and therefore may be used

to filter the whole collection into a smaller set before performing the more costly histogram comparisons (Hafner et al , 1995)

- **MPEG7 Colour Layout** The MPEG7 Colour Layout description can be used to describe the spatial layout of colours in an image or region The average colour for each cell in a 8x8 spatial grid is calculated and the DCT transform is performed (see next section on texture features, p 59, for a description of DCT) Similar to the JPEG standard, the most significant DCT coefficients, i e the DC coefficient and some of the low frequency AC coefficients are stored The Y, Cb and Cr channels of the YCbCr are transformed separately and normally more AC coefficients are kept for the Y luminance channel in comparison to the Cb and Cr colour channels

- **Dominant Colours** An extension to average colour description is to represent the image or region with a set of dominant colours

- **MPEG7 Dominant Colour** The MPEG7 Dominant Colour descriptor compactly describes an image or an arbitrary region with a set of between 1 and 8 colours Each dominant colour's percentage covering of the image and its colour variance (a single bit indicating low-variance versus high-variance) is calculated The Dominant Colour descriptor can be associated with a colourspace and colour quantisation information that defines the discrete colour codes in the calculation of the dominant colours It also contains a single spatial coherence value on a 31-point scale for all the dominant colours Dominant colour is therefore a limited but compact representation of the colour in a video segment, image or region It is more suitable for representing colours of objects or image regions where a limited number of colours may be sufficient than whole images (Ojala et al , 2002)

- **Colour Sets** Colour sets result from a bmarisation of the colour histogram's bins In VisualSEEk colour sets are generated from a 166 bin HSV histogram for whole images and their subregions (Smith and Chang, 1996a) Quadratic and other cross-bin similarity functions are more efficient when calculated on the binary representation of the histogram Similar to dominant colours and colour moments, this representation is more appropriate for subregions that have been segmented using an appropriate colour homogeneity principle than for representing whole images

- **Colour Histograms** Histograms are the defacto standard for representing colour information in an image and can model the joint distribution of the three colour channels or the marginal distribution of colour channels separately The attraction of colour histograms is that they are somewhat rotational and zooming invariant and can characterise the colour of the whole image across the full colourspace Quantisation of the colourspace is used to reduce the space requirements of the histogram The colour histogram colour feature is represented in many different ways for current video retrieval systems As a global image feature it can be

represented as an HSV histogram (Smith et al , 2002, Hauptmann et al , 2003), as a LAB histogram (Worring et al , 2004, Snoek et al , 2005), a RGB histogram (Wu et al , 2004) or an HMMD Colour Histogram (Pickering et al , 2003) Independence is sometimes assumed between the different dimensions of the histogram (marginal distributions of colour channels) leading to compact representation for retrieval Local colour features such as a grid-based LAB colour histogram (Wu et al , 2004), a grid-based HSV histogram (Smith et al , 2002) and a grid-based Munsell and RGB histogram (Hauptmann et al , 2002) are also used for video retrieval

- **MPEG7 Scalable Colour** The MPEG7 Scalable Colour descriptor is an Haar encoded HSV histogram description of the colour of an image or region The colourspace and quantisation is constrained to the HSV colourspace with a uniform 256 bins (16 hue x 4 saturation x 4 brightness) and is encoded using the Haar transform to reduce the space required to store the histogram as well as providing a means to encode only the most significant information Scalable colour is a compact representation of the global distribution of colour within an image or region

- **Colour Coherence Vectors** Colour coherence vectors are a refinement of colour histograms that stores the number of coherent versus incoherent pixels for each quantised colour (Pass et al , 1996) A colour is defined as coherent if it is part of a larger similarly-coloured region

- **MPEG7 Colour Structure** The MPEG7 Colour structure description is an histogram-like feature that describes the local structure of colour within the whole image It is calculated by sliding an 8x8 square structuring window over the image and incrementing the counts for each colour present in the window The HMMD colourspace is used with either 255, 128, 64 or 32 quantisation levels The Colour Structure Descriptor is used in (Pickering et al , 2003) for TRECVid video shot retrieval

- **Colour Correlograms** Another alternative to the histogram that incorporates spatial information is the Colour Correlogram, which describes the global distribution of local spatial correlations of colours based on spatial distances between colour pixels (Huang et al , 1997) The Colour Correlogram can be considered as partly a texture description as well as a colour description A colour correlogram defines for each colour pair $< c_i, c_j >$ the probability of finding a pixel of colour c_j at a distance k from a pixel of colour c_i in the image A set of distances are used to generate the correlogram Formally, the correlogram entry for the colour pair $< c_i, c_j >$ at distance k in image $\mathcal{I}$ is defined as

$$\text{Correlogram}(c_i, c_j, k) = \mathbb{P}(\mathcal{I}(p_2) = c_j | \mathcal{I}(p_1) = c_i \wedge \mathcal{I}(p_1, p_2) = k) \qquad (67)$$

where p_1 and p_2 are any two pixel indexes, $\mathcal{I}(p)$ defines the pixel colour index and

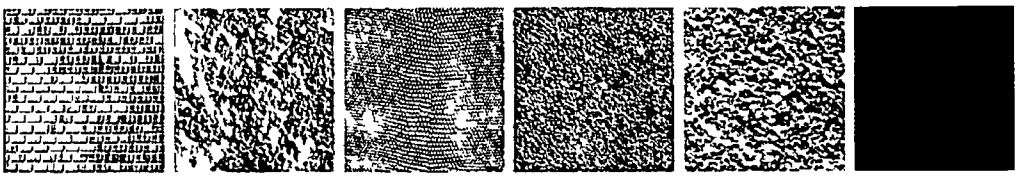


Figure 10 Texture examples from the Brodatz dataset (Brodatz, 1966)

$\mathcal{I}(p_1, p_2)$ defines the distance between pixels A geometric based distance function is usually selected such as Manhattan, Euclidean, or even L_∞ The space requirements for a correlogram with N colour values and D distances is $O(N^2D)$ Due to the large space requirements, it is common to use a simplification of the Colour Correlogram called the *autocorrelogram* that models only the spatial correlation between identical colours (i e autocorrelogram(c,k) = correlogram(c,c,k)) Darwish et al (2002) use a HSV Colour Correlogram for the TRECVid search task

Ma and Zhang (1998) compared colour histograms, moments, correlograms and coherence vectors and found that correlograms performed best for image retrieval Ojala et al (2002) compared the Colour Layout, Colour Structure, Dominant Colour and Scalable Colour MPEG7 descriptors with the HSV autocorrelogram for the retrieval of semantic image categories Eight categories covering 822 images in total were manually created from a collection of 2445 images In their experiment they let each of the 822 images serve as a query They concluded that local spatial organisation of colours is important for retrieval – the MPEG7 Colour Structure description performed best while the non-spatial MPEG7 Dominant Colour performed worst Overall the non-spatial Scalable Colour (global HSV Colour Histogram) was second best, performing better than the third best feature, HSV Colour Autocorrelogram, at all recall levels The Colour Layout performed second worst at high-precision but was best at high-recall This experiment may have limited applicability to video retrieval due to its small size and the nature of the image collection and topics, which were based on image classification as opposed to video retrieval

3 2 4 2 Texture Descriptions

In contrast to colour, which can be defined independently for single pixels, texture is considered a property of a local spatial neighbourhood within the image There is no exact definition of texture and it is more easily recognised than defined (see Figure 10 for texture examples from the Brodatz texture dataset) We will loosely define image texture as patterns within a local spatial neighbourhood that can occur at different scales within the image

There are many competing types of texture representation such as statistical, geometrical and signal processing Statistical methods include statistics extracted from a co-occurrence matrix of gray levels at different distances (Haralick et al , 1973) Geometrical models of texture represent texture as repeated geometrical shapes and we do not discuss these further as they are not useful for the general video retrieval task Signal processing methods represent texture based on discrete image transforms, such as Gabor filters and DCTs, of the image or image tiles Gabor wavelets are currently one of the more popular methods of extracting texture for the video retrieval task (Hauptmann et al , 2003) The MPEG7 standard contains three texture descriptions, Homogenous Texture, Texture Browsing and Edge Histogram While the edge feature is defined as a texture description in the MPEG7 standard, which it partly is, we will describe it in the next section on extracting shape descriptions We will now describe some of the texture representations used in image and video retrieval

- **Co-occurrence statistics** Co-occurrence statistics of gray levels are one of the traditional methods of characterising the texture of an image A gray-level co-occurrence matrix is first generated for pixels separated at a fixed vector (angle and distance) across the image and statistics of this co-occurrence matrix are used to represent the texture of the image Haralick et al (1973) proposed many such statistics including entropy, energy, contrast, homogeneity, sum mean, variance, correlation, maximum probability, inverse difference moment and cluster tendency Five of these statistics namely energy, entropy, contrast, homogeneity and correlation are more popular than the others in image retrieval and these are defined as

$$Energy = \sum_i \sum_j \mathbb{P}(i,j)^2 \tag{68}$$

$$Entropy = -\sum_i \sum_j \mathbb{P}(i,j) \log \mathbb{P}(i,j) \tag{69}$$

$$Contrast = \sum_i \sum_j (i-j)^2 \mathbb{P}(i,j) \tag{70}$$

$$Homogeneity = \sum_i \sum_j \frac{\mathbb{P}(i,j)}{1+|i-j|} \tag{71}$$

$$Correlation = \sum_i \sum_j \frac{(i-\mu_i)(j-\mu_j)\mathbb{P}(i,j)}{\sigma_i * \sigma_j} \tag{72}$$

where $\mathbb{P}(i,j)$ is the probability of gray levels i and j being separated spatially by the given vector (i e a given angle and distance) that defines the co-occurrence matrix, μ_i and μ_j are the respective mean gray-levels and σ_i, σ_j are the respective standard deviations In some representations, the co-occurrence matrix is calculated for a fixed pixel distance without a specified angle Howarth and Ruger

(a) 1D Gabor filter (b) 2D Gabor filter

Figure 11: Example of a (a) 1D Gabor filter and (b) 2D Gabor filter

(2004) compared energy, entropy, contrast and homogeneity statistics for different parameterisations of the co-occurrence matrix on the Corel image dataset and found that the homogeneity statistic performs best for this image retrieval task.

- **Tamura Texture** Tamura et al. (1978) defined five features, three of which, namely coarseness, contrast and directionality, are popular in image and video retrieval (Equitz, 1993). Coarseness defines the largest size of the repeating texture. Contrast defines the gray-level range of the texture. Directionality quantifies the dominance of a specific direction within the texture. In both MARS (Ortega et al., 1997) and the texture experiments of Howarth and Rüger (2004), they create a Tamura image which is the joint distribution of these three features at different pixel locations and they represent this feature using both marginal and multidimensional histograms.

- **MPEG7 Texture Browsing** The MPEG7 Texture Browsing description characterises texture in terms of regularity, coarseness and directionality. Regularity is defined on a four point scale – irregular, slightly irregular, regular and highly regular. Direction and Scale can be specified either once or twice if there are two dominant aspects to the texture. The texture's scale is on a 4 point scale – fine, medium, coarse and very coarse. The texture's direction can be specified as non-directional or on 30 degree increments. This highly compact texture description is more appropriate for browsing textured images (hence its name) than for supporting content-based querying.

- **Gabor** The Fourier transform represents an image in terms of spatial frequencies but is inappropriate for texture representation because the spatial frequencies depend on every pixel in the image and are therefore not capable of representing texture that is localised at different spatial scales. A Gabor filter is a Gaussian envelope modulated by a sinusoidal plane wave (see Figure 11) and a bank of such filters at different scales and orientations can be applied to an image to extract a texture description. The scale of the Gabor filter (standard deviation of the Gaussian envelope) limits the spatial extent of the filter thereby localising texture extraction to a specific size within the image. The orientations of the Gabor filter defines the direction of the spatial frequencies to extract. The Gabor filter is

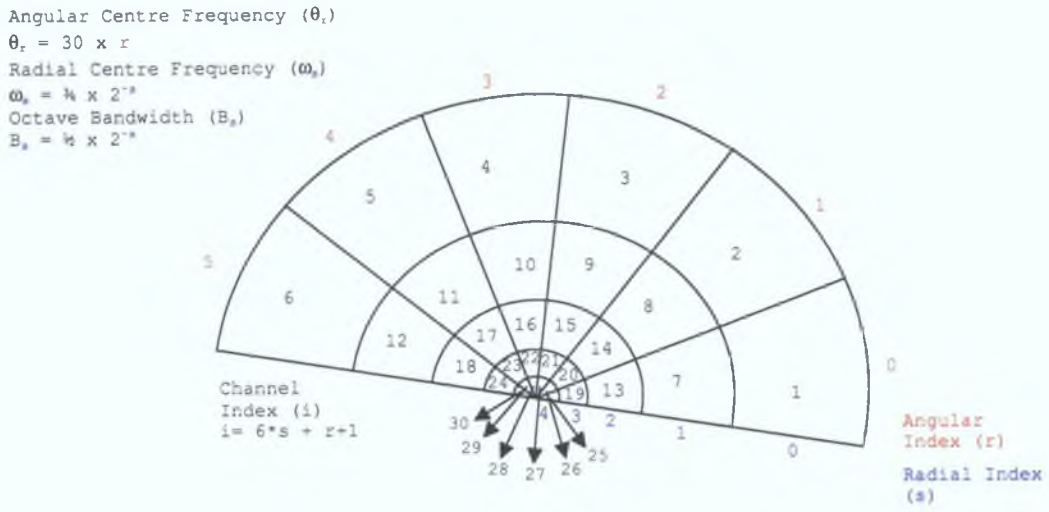


Figure 12: Frequency layout for the MPEG7 Homogeneous Texture descriptor.

defined as

$$G(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)} e^{j2\pi Wx} \tag{73}$$

where σ_x and σ_y are the standard deviation of the Gaussian envelope and W is the modulation frequency. Different scales and orientations of this filter are achieved by appropriate dilation and rotation of this function. Gabor filter banks improve on texture representation over the related Fourier transform due to the localisation of the spatial extent of the extracted frequencies. A Gabor wavelet is a Gabor filter bank containing a quasi-orthogonal subset of Gabor filters. Howarth and Rüger (2004) compared different Gabor filter banks on the Corel image dataset and found that 7x7 tiling (i.e. breaking the image into 49 image tiles) with a Gabor filter bank with 2 scales and 4 orientations performed best. They suggest that a higher number of scales may be more effected by noise at the coarser scales. For TRECVid video search Hauptmann et al. (2003) broke each shot's keyframe into a 3x3 rectangular grid and used Gabor filters for 6 angles at a single scale and their central and second order moments were indexed.

- **MPEG7 Homogenous Texture** The MPEG7 homogeneous texture description represents the texture by the mean and standard deviation of the image intensity and the mean and optionally the deviation of the energy in 30 Gabor frequency channels. The Gabor frequency channels are laid out in 6 orientations of 30 degrees and 5 radial centre frequencies spaced in an octave scale (see Figure 12). The mean energy of each Gabor channel is defined as the log-scaled sum of the square of the Gabor-filtered Fourier transform coefficients of the image, while the energy deviation is the log-scaled standard deviation of the square of Gabor-filtered Fourier transform coefficients. The interested reader may find the formal definitions of these within the MPEG7 standard (MPEG7 Committee, 2002). The

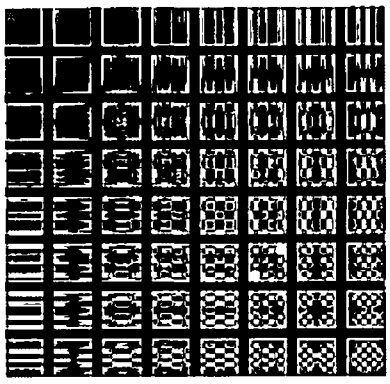


Figure 13 DCT basis functions

average and standard deviation of image intensity and Gabor energies are each uniformly quantised into 256 values

- **DCT** The DCT transform (Discrete Cosine Transform) is an alternative signal processing method to Gabor filters for texture representation This feature is similar to the MPEG7 Colour Layout feature but is normally extracted for small rectangular blocks (8x8 pixels) in the whole image instead of being extracted for the whole image as in the MPEG7 Colour Layout descriptor The DCT transform is similar to the Fourier transform but is computationally simpler and only models the real part (magnitude) of the Fourier transform The DCT transform $F(u,v)$ of an 8x8 image block $f(x,y)$ is defined as

$$F(u,v) = \frac{C_u}{2}\frac{C_v}{2}\sum_{y=0}^{7}\sum_{x=0}^{7} f(x,y)\cos\left(\frac{(2x+1)u\pi}{16}\right)\cos\left(\frac{(2y+1)v\pi}{16}\right), \quad (74)$$

where

$$C_u = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } u = 0, \\ 1 & \text{if } u > 0 \end{cases}, \; C_v = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } u = 0, \\ 1 & \text{if } u > 0 \end{cases}$$

The distribution of the DCT coefficients in the rectangular blocks characterise the spatially localised texture within the image (see Figure 13 for visualisation of each coefficient's basis function) Vasconcelos and Lippman (1998) used the DCT feature as a joint colour and texture feature for image texture retrieval and modelled its distribution using a Gaussian Mixture Model (GMM) The GMM model is learned using the EM algorithm (Dempster et al , 1977) from the distribution of the image blocks' DCT coefficients A diagonal covariance matrix may be assumed for the GMM model to increase the speed of the iterative EM algorithm when indexing the image collection Baan et al (2002) combined this generative model, a Gaussian Mixture Model of multi-spectral DCT coefficients of keyframe images, with an hierarchical language model of ASR text and applied it to the TRECVid video search task Similar to the MPEG7 Colour Layout description, only the most significant DCT coefficients are modelled and normally more AC

coefficients are kept for the Y luminance channel in comparison to the Cb and Cr colour channels

Howarth and Ruger (2004) compared three types of texture descriptions, namely co-occurrence statistics, Tamura features and Gabor filters for retrieval on the TRECVid 2003 and Corel datasets Overall, the Gabor features performed best and the co-occurrence statistics performed strongly They reported a relatively high Mean Average Precision (MAP) of 0 0431 on the TRECVid 2003 task for visual-only query-by-example searching when they combined the results of their Gabor feature with their HSV colour feature

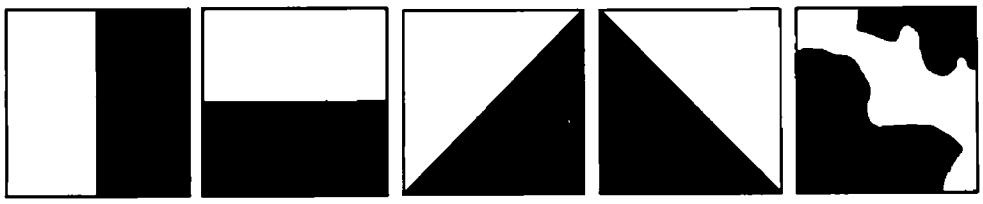
3 2 4 3 Shape Descriptions

The shape of an object can be thought of as its silhouette Unfortunately, current automatic image segmentation techniques for general video content are not robust enough to segment real-world objects from images and video segments Instead homogeneous regions which are spatially connected and share a common colour or texture are usually extracted

An edge is an abrupt change in image intensity and can have both an associated orientation (direction) and magnitude The distribution of edges in an image provides limited support for shape retrieval Whereas colour is directly extracted from an image or video, edges are recognised using edge detectors such as based on the Laplacian, Prewitt, Sobel, Robinson, or Kirsch operators (Sonka et al , 1998) The Canny edge detector (Canny, 1986) is one of the more effective standard edge detectors and has been used for trademark retrieval (Jain and Vailaya, 1998) The Canny edge detector is a three stage process First, the image is smoothed using a Gaussian convolution to remove noise and control the level of detail in the image Second, a 2D first derivative operator is applied to the image Finally, non-maximal suppression is applied by tracking the tops of ridges of the edge gradient and setting adjacent non-edge pixels to zero Edges are first identified by being above a high threshold and are tracked until the gradient is below a lower threshold - a process known as hysteresis

The edge representations used in visual retrieval include edge histograms, edge co-occurrence matrices and edge correlograms The MPEG7 standard supports the description of the edges within an image using an edge orientation histogram and also supports more direct encoding of the shape of regions and objects with the MPEG7 Region Shape, MPEG7 Contour Shape and MPEG7 Shape 3D descriptors We will now describe these shape features

- **Edge Histogram**



(a) Vertical (b) Horizontal (c) 45 degree (d) 135 degree (e) Non-directional

Figure 14 MPEG7 Edges

An edge histogram can represent the distribution of edge orientations and edge strengths within an image It is usual to model the non-occurrence of edges in the histogram and also to ignore the strength of edges in the histogram representation (Hauptmann et al , 2004) recognised edges using the Canny edge detector and simply represented the edge by its quantised direction, while Smith et al (2002) used the Sobel edge filter with 8 angles and 8 magnitudes for video retrieval Wu et al (2004) segment the TRECVid keyframes into a 4x4 grid before extracting an edge direction histogram for the TRECVid search task

- **MPEG7 Edge Histogram** The MPEG7 Edge histogram counts the number of 5 different edge types in 16 rectangular regions (4x4 grid) of an image The edges are vertical, horizontal, 45 degrees, 135 degrees and non-directional (see Figure 14) The histogram bins are normalised by the number of pixels in the source image and therefore indirectly represents the number of non-edge pixels The histogram bins are non-linearly quantised into 8 values each Won et al (2002) evaluate the MPEG7 Edge histogram feature using 51 image queries on an image set of 11639 images from the MPEG7 Core Experiment They find that using global and semi-global histograms extracted from the local edge direction histograms in the similarity function improves image matching performance over just using the local edge histograms

- **Edge Co-Occurrence Matrix** Similar to colour, the edge histogram feature can be extended to model the co-occurrence of edges types The edge co-occurrence matrix models the distribution of edge pairs in their local neighbourhood throughout the whole image Brandt et al (2002) suggested this feature with 8 edge directions (8x8 co-occurrence matrix) and evaluated it for recognising different shapes (aircraft, buildings, faces)

- **Edge Correlogram** The correlogram feature, which we previously discussed for modelling colour, can also be applied to the global distribution of local edge correlations This model was further applied to a sequence of images in which it was referred to as the Temporal Gradient Correlogram (Rautiainen, Seppanen, Penttila and Peltola, 2003) For this type of Gradient Correlograms, the edges were detected using the Prewitt edge detector for four edge orientations and an

auto-correlogram was generated for 4 spatial pixel distances (1, 3, 5, 7) measured using the L_∞ norm

- **MPEG7 Region Shape** The MPEG7 Region Shape feature describes the shape of an arbitrary object, which does not need be fully connected and can contain holes The shape of the region is stored as 35 normalised and quantised magnitudes of ART (Angular Radial Transform) coefficients The 35 magnitudes consist of the shape's response to 12 angular and three radial functions The interested reader may find more information on this and other MPEG7 shape descriptors in (MPEG7 Committee, 2002)

- **MPEG7 Contour Shape** The MPEG7 Contour Shape descriptions represents the shape of a closed contour of a 2D image region The shape is represented in a Curvature Scale Space and is smoothed with a filter until it becomes convex The MPEG7 Contour Shape description specifies the circularity and eccentricity of the original shape and of the shape after termination of the filter The highest peak and optionally up to 62 less prominent peaks of the Curvature Scale Space image are stored

- **MPEG7 Shape 3D** The MPEG7 Shape description allows 3-dimensional shapes to be described with a 3D mesh model

Brandt et al (2002) compared indirect shape features such as the histogram of 8 edge directions, the co-occurrence matrix of edge directions and the decimated Fourier transform of the edge image that characterise the shape of objects within an image but do not require image segmentation They evaluated these 'shape' features on the task of detecting images containing aircraft, buildings and faces and found that their best results were given by the decimated magnitude Fourier spectrum of the edge image and that the edge co-occurrence matrix of edge directions also achieved good results They suggest with some support from their experiment's results that invariance of features to all affine transformations (rotation, translation and scaling) has a negative effect on retrieval/classification performance for general image collections

To truly model the shape of 3D objects projected onto 2D representations is far more difficult than we have discussed here Edge-based features are an indirect representation of the shape of objects within a scene and in practice may strongly characterise the texture of the object more so than its outline silhouette Indeed the 2D silhouette of an object varies considerably depending on the angle that the 3D object is viewed from The shape of an object also varies as the object or parts of it moves (e g a static camera view of a hand varies considerably as the person opens and closes his fist) We can also view the shape of an object as being characterised by the shape of its parts but it is extremely difficult in an ad hoc retrieval task to automatically extract the component shapes and their inter-relations from a small set of query images The statistical analysis

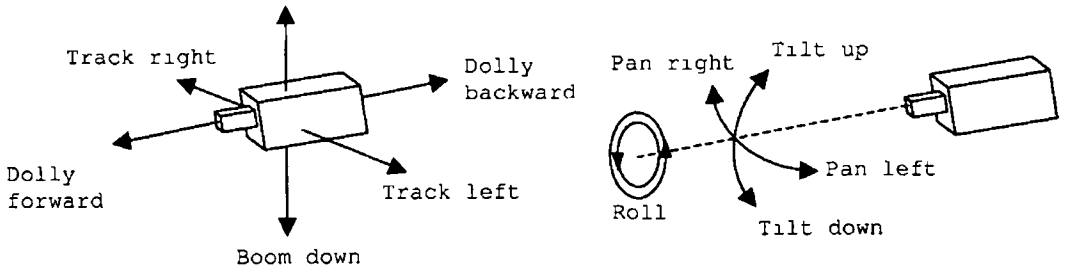


Figure 15 Illustration of MPEG7 Camera motion – reproduced from (MPEG7 Committee, 2002)

of shapes, such as with Active Shape Models (Stegmann and Gomez, 2002), provides a richer representation of the shape of objects but requires larger amounts of annotated training data than is usually available in ad hoc retrieval scenarios These more advanced shape models may be usefully applied at the video indexing stage in order to identify object classes (e g vehicles, people, buildings) that would be useful to users at retrieval time

3 2 4 4 Motion Descriptions

The temporal dimension distinguishes videos from still images The motion of the video camera and objects within a video shot are sometimes important factors for users searching for video in order to reuse video segments in a new production Video retrieval is an interactive process of querying and browsing and in many video retrieval systems motion is not supported in the specification of visual queries Users in these systems are reliant on browsing and playback of their video retrieval results in order to discern their motion Representation in the user interface of the constituent motion, video mosaicing (Irani et al , 1996) and video skims of the retrieved video segments (Smith and Kanade, 1995) provide browsing structures that may support users in quickly distinguishing between video retrieval results in terms of their constituent motion

The MPEG7 standard supports the representation of camera motion and three object motion descriptors namely MPEG7 Motion trajectory, MPEG7 Parametric motion and MPEG7 Motion activity We will now describe these motion features

- **MPEG7 Camera motion** The MPEG7 Camera motion descriptor characterises the 3-dimensional motion of the video camera Supported camera motions include fixed, panning, tracking, tilting, booming, zooming, dollying and rolling (see Figure 15) The described shots can be broken up into sub-shots with different camera motion defined for each of these temporal segments The atomic camera motion types may be combined to describe the effective motion of the camera in a particular sub-shot The MPEG7 Camera motion feature also includes the specification

of the speed of induced image motion and the focus of expansion or contraction in cases where the camera is dollying or zooming The focus point is the point in the image where the camera is centrally focussed when zooming or dollying

- **MPEG7 Motion trajectory** The MPEG7 Motion trajectory descriptor allows the 2D and 3D trajectory of moving regions within video segments to be specified using key-points The motion trajectory between these key-points is interpolated using a velocity model that may optionally include acceleration

- **MPEG7 Parametric motion** The MPEG7 Parametric motion descriptor supports the description of global and object motion using 2D parametric geometric models such as translation, rotation, affine, perspective and quadratic models The parametric motion descriptor may be associated with an MPEG7 Moving Region and motion is measured in terms of pixel displacement

- **MPEG7 Motion activity** The MPEG7 Motion Activity description consists of five features, namely intensity, dominant direction, spatial distribution, spatial localisation and temporal distribution, that may be calculated from the motion vectors within a video sequence Motion intensity is measured on a five point scale, while the dominant direction of the motion activity is quantised into 8 directions The spatial distribution of motion activity is measured in terms of the number of short, medium, and long runs of zero motion within the video and indicates whether the motion activity is spread across multiple spatial locations or is localised to a single region The spatial localisation feature models the spatially localised motion intensities over the duration of the video segment The video is broken into either 2x2, 4x4, 8x8, or 16x16 pixel sized blocks and this description quantises the average intensity over the whole video segment of each individual sub-image block into 8 values The temporal distribution feature characterises the activity pattern over the duration of the video sequence using a 5-bin histogram of the intensity values of motion within the video sequence

It is envisaged in the MPEG7 standard that video cameras could directly record and encode their camera motion as metadata associated with the captured video At present however, the computation of optical flow fields forms the basis of many current techniques for determining camera and object motion (Akutsu and Tonomura, 1994) Optical flow is an approximation of image velocity on the 2-dimensional image plane – see (Beauchemin and Barron, 1995) for a review of optical flow methods The principle for using optical flow to detect camera and object motion is the assumption that dominant motion of the video's optical flow field can be associated with the camera motion, while localised optical flow that differs from this can be associated with objects Determining camera and object motion automatically from video is currently a challenging research topic

Camera and object motion has been integrated into several video retrieval systems, e g SWIM (Zhang et al , 1997), VideoQ (Chang et al , 1997) and QBIC (Flickner et al , 1995), but there has been no systematic empirical evaluation of different object motion techniques for general video retrieval It is probable that motion is a feature that must be selectively chosen in video retrieval when performing query-by-example searching The VideoQ system is one of the earliest video retrieval systems to incorporate automatic object motion detection and to support motion-based retrieval It automatically recognises object motion from analysis of the optical flow, smoothes the recognised trajectories using Mallat wavelets (Mallat, 1991), segments the trajectories at their points of maximum acceleration and indexes each sub-trajectory by its order, arc length, edge points and its acceleration and velocity in the x and y directions (Chen and Chang, 2000) Matching a trajectory at retrieval time is achieved by first matching sub-trajectories using Mahalanobis distance (see Section 3 3 4) of all attributes except order and then combining results based on the order attribute

3 2 4 5 Face Descriptions

Face detection and recognition functionality potentially provides a useful semantic index to aid searching and browsing of video collections Unlike standard face matching from frontal passport photographs, face matching from general video content is far more challenging since faces can appear in many different lighting conditions, sizes, poses, facial expressions and poorer image resolutions Most research into face recognition concerns still images with face recognition in video sequences receiving less attention – see (Turk and Pentland, 1991, Shio and Sklansky, 1991, McKenna et al , 1997, Torres and Vila, 2001) for research on face recognition using video sequences

The MPEG7 Face Recognition descriptor, which is based on the popular Eigenface approach (Turk and Pentland, 1991), provides a standard method for describing faces Other approaches to face detection and recognition identify parts of the face such as eyes, nose, mouth and ears and model their relationship with for example a graph (Wiskott et al , 1997) Comprehensive surveys on different techniques for face detection and recognition can be found in (Chellappa et al , 1995, Yang et al , 2002) We will briefly describe the Eigenface and related MPEG7 Face Recognition descriptor

- **Eigenfaces** In the Eigenface approach to face detection, face images are scaled to a common size and principal component analysis (PCA) is performed on the group of images (Turk and Pentland, 1991) PCA produces a set of orthogonal Eigenvectors that are ordered by their discrimination power as measured by their respective eigenvalue A fixed number of the top Eigenvectors are retained providing a face discrimination space of reduced dimensionality that retains much of the discrimination power for the original set of face images Face matching is

performed by projecting a test face image into this Eigenface space and measuring the distance between it and existing indexed faces using typically either Euclidean or Mahalanobis distance (Lorente and Torres, 1999)

- **MPEG7 Face Recognition** The MPEG7 Face Recognition descriptor is a specific instance of the Eigenface approach (i e with a fixed image size and prespecified basis vectors) The face is first extracted from the video frame and represented as luminance values within a normalised 46x56 face image This image vector is projected onto a different basis and then normalised and clipped to produce a face vector of 48 values, which is further quantised to a 5-bit range for each value

Automatically indexing faces in video requires two stages of video processing - face detection and face recognition Face detection is the first stage and involves identifying regions of video frames that contain faces Candidate face regions can be identified by locating skin coloured pixels in the image that form a connected region in the shape of a face and an Eigenface projection of these regions can be used to verify whether they actually contain a face (Czirjek et al , 2003) Video retrieval systems generally stop at detection and simply index by the number and possibly size of the faces found in each keyframe due to the previously mentioned difficultly of face recognition techniques in general video collections

Chen and Hauptmann (2004) performed face recognition using Eigenfaces for a subset of five TRECVid 2003 person-finding topics They supplemented these TRECVid person queries with extra varied face images of the target person and also indexed all faces in all I-Frames of the shots in order to reduce the problem of faces appearing in different poses I-Frames are video frames that are independently encoded in the MPEG1 video stream with relatively high quality compared to other frames and occur regularly in the video stream, typically around one every 12 frames The inclusion of face recognition increased the mean average precision by 8 5% (MAP 0 420 compared to 0 387) for the set of five tested person finding topics While these results are for a very small set of topics and are not directly applicable to person queries with few visual examples, they indicate that there is merit in integrating current face recognition techniques into general video retrieval systems

3 2 4 6 Spatio-Temporal Descriptions

There are not many spatio-temporal feature models beyond object and camera motion models currently in use in video retrieval In fact the temporal dimension is generally under-utilised for video retrieval with nearly all video retrieval systems converting video retrieval into a form of still image retrieval by the process of shot segmentation and keyframe extraction

The potential advantage of using features extracted from video sequences instead of a single keyframe is that the extracted description may be more likely to represent the shot and is less dependent on the keyframe selection method If the feature varies significantly throughout the shot then extracting multiple keyframes or multiple key video sequences may be more effective than trying to represent the shot by a single feature representation averaged over the shot

The previously described spatial features can be extended to represent spatio-temporal image sequences simply by averaging or aggregating the representation over multiple images There are a few extensions of colour, edge and texture descriptions that have been applied to video sequences as opposed to keyframes for the video retrieval task, which we now describe

- **MPEG7 GoF/GoP Colour** The MPEG7 GoF/GoP (Group of Frames/Group of Pictures) Colour description uses the MPEG7 Scalable Colour descriptor to describe the average, median, or intersection HSV histogram of the colour of a group of frames or pictures It does not explicitly model the temporal dimension but allows an MPEG7 Scalable Colour descriptor (Haar encoded HSV colour histogram) to be associated with a spatio-temporal video segment

- **Temporal Colour Correlogram** The Temporal Colour Correlogram, an extension of the HSV Colour Correlogram, models the spatial relationship of colours in a video sequence with co-occurrence statistics (Darwish et al , 2002, Rautiainen and Doermann, 2002) The feature is extracted from 20 frames evenly distributed within the video shot and the auto-correlogram was generated for a set of colours at 4 distances (1, 3, 5, 7 pixels) This model also does not explicitly model the temporal dimension of the colour in the video sequence

- **Temporal Gradient Correlogram** Rautiainen, Penttila, Pietarila, Noponen, Hosio, Koskela, Makela, Peltola, Liu, Ojala and Seppanen (2004) extract a gradient (edge) correlogram from a series of frames from the same shot Similar to the Temporal Colour Correlogram, this feature was extracted from 20 evenly distributed frames and did not explicitly model the temporal dimension in the representation The correlogram was generated for 4 edge orientations at 4 distances (1, 3, 5, 7 pixels)

- **Dynamic GMM DCT** Westerveld et al (2004) extend their GMM generative model for spatial DCT-based joint colour and texture feature representations to include the temporal dimension This was achieved by adding the time dimension to the feature vector and sampling 29 frames, 14 frames before and after the keyframe This model is further extended in Ianeva et al (2005) by incorporating a full co-variance model instead of a diagonal covariance model in the Gaussian Mixture Model The full co-variance model frees the components of the GMM

mixture from having to be aligned with the feature axis, which is the case for a diagonal co-variance model

Comparisons of these spatio-temporal representations with their related spatial-only representation have not shown significant improvements in retrieval effectiveness Ianeva et al (2004) compared the spatial-only GMM DCT representation with the dynamic spatio-temporal GMM DCT representation and report the same MAP of 0 022 but a higher precision at document cut-off 10 for the dynamic model (0 096 compared to 0 076) on the TRECVid 2003 collection In TRECVid 2004 the dynamic model with full covariance was compared with the spatial-only model (Ianeva et al , 2005) and the dynamic model achieved a slightly higher MAP of 0 010 compared to 0 008 but lower precision at document cut-off 10, 15, 20 and 30 Temporal colour correlograms are compared with colour correlograms for the TRECVid 2001 search tasks in (Rautiainen and Doermann, 2002, Darwish et al , 2002) and were found to be many times better than colour correlograms though this is maybe the result of a poor keyframe selection method since their keyframe selection method chose the first frame of the shot, which is perhaps a poor choice (middle would be safer) as this could be affected by gradual transitions

The spatio-temporal features that we have described have very restricted representation, if any, of the temporal patterns of the visual features Temporal textures are an alternative spatio-temporal feature that seek to represent textures that are strongly defined in the temporal dimension such as flowing water, flames and foliage (Nelson and Polana, 1992) The motion of crowds of people as viewed from a distance may also be categorised as a temporal texture (Bouthemy and Fable, 1998) Temporal textures may be useful for content-based search but as yet this type of feature has not been explored for the TRECVid general video search task

3 2 5 Audio Descriptions

In TV news, documentary and general television video collections the primary information in the audio channel is speech, which as previously mentioned has been shown to be the most effective single feature for the visually oriented TRECVid search tasks Speech can be recognised using an automatic speech recogniser and thereby transformed into a textual description Most further use of the audio stream in video retrieval is in speaker segmentation and the detection of high-level features such as speech, instrumental sound and monologue segments

There is a limited amount of research that has investigated the use of audio features directly in content-based video search systems In TRECVid 2002 Adams et al (2003) transformed the uncompressed audio stream into 24 Mel-Frequency Cepstral Coefficients

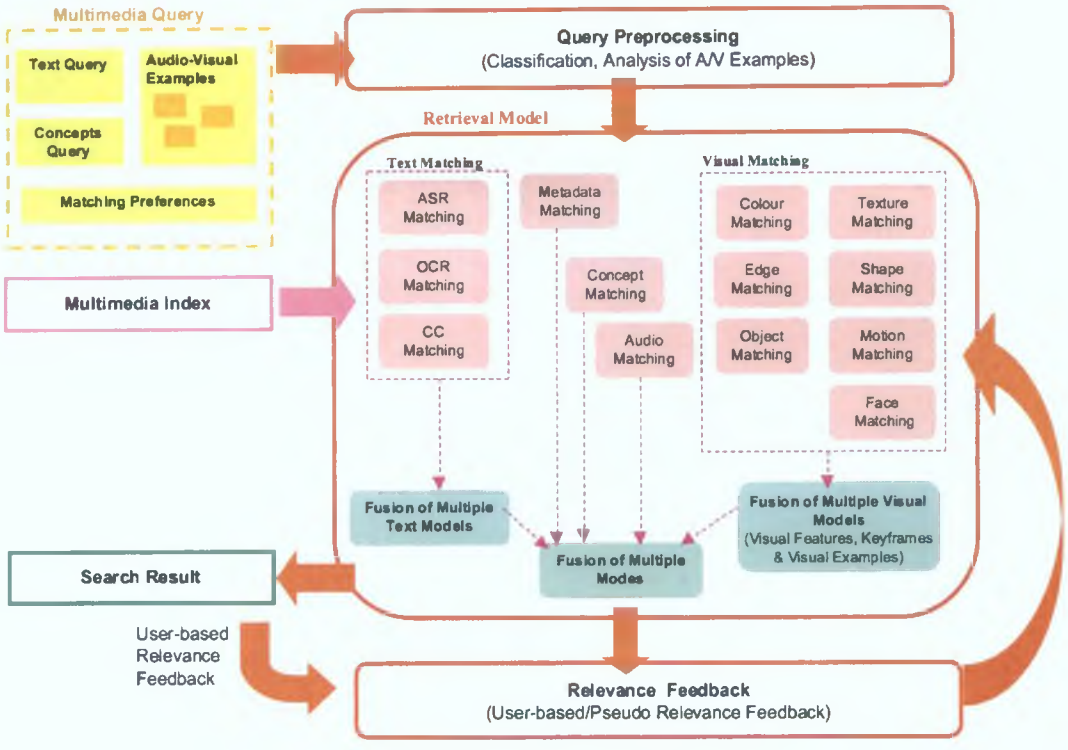


Figure 16: Overview of video retrieval system.

(MFCC) which describe the acoustic energy in different frequency bands and made this feature available to all TRECVid participants in the manual and interactive search experiments. In TRECVid 2004 Ianeva et al. (2005) modelled the distribution of MFCCs in shots using their generative Gaussian Mixture Model approach but found that their GMM MFCCs feature did not perform well and in fact decreased performance when combined with other features.

3.3 Video Retrieval Models

In the previous section we described the features that can be indexed to support video retrieval while in this section we will discuss the video retrieval process itself. We illustrate the components of a state-of-the-art video retrieval system in Figure 16 for handling a multimedia query consisting of a text description, high-level concepts and multiple audiovisual examples such as images and video clips. There are three main stages to handling a multimedia query: query preprocessing, query matching and relevance feedback.

Retrieval systems may first preprocess the query by extracting any required features from audiovisual examples that are used in the query but are not part of the indexed collection. The query may also be automatically classified into predefined query categories that define the initial set of features and fusion settings for matching the query and

retrieval documents An example set of query-classes might classify a query as a request for named persons, named objects, general objects and scenes (Hauptmann et al , 2004, Yan et al , 2004) Alternatively, the user may have pre-specified the query-class or some matching preferences when composing the query

The next stage is to generate an initial ranking of video retrieval units in response to the query This involves first separately matching each query feature and query example with the retrieval units and then fusing these separate scores into a single result list In most retrieval experiments the video retrieval units are generally camera shots though in practice users may prefer retrieval units such as videos, stories, or sequences of shots for some types of queries The text query may be matched against multiple text indexes such as automatic speech recognition (ASR) text, closed captions (CC) text and video optical character recognition (video OCR) text The query's visual examples may be matched using low-level visual features such as colour, edge, texture, motion as well as higher-level features such as segmented objects or faces The query may also be matched to retrieval units based on semantic concepts, video metadata and other audio features All these feature matching results associate a retrieval score to each retrieval unit These scores are fused into a single result for example (the order of fusion tasks is not essential) by combining all text scores into a single result, combining visual feature scores into a single result for each query example, then combining all query example scores into a single visual result and finally by combining all multimodal scores such as for text, visual, audio, concepts and metadata

After creating the initial ranking, it may be beneficial to perform pseudo-relevance feedback, which usually assumes that the top ranked documents are relevant and/or the low ranking documents are non-relevant Pseudo-relevance feedback facilitates the reweighting and expansion of the components of the query For example the query text may be expanded or high-level concepts that provide corroborative information with the initial results may be added to the query using text or high level concepts taken automatically from the underlying top-ranked video shots The retrieval results are presented to the user whom may also provide more reliable relevance information on which the retrieval system may base more effective relevance feedback upon

In the following sections we will describe the key processes in a video retrieval system query pre-processing, text matching, concept matching, visual matching, fusion and relevance feedback

3 3 1 Query Preprocessing

A recent advancement in query pre-processing is the automatic classification of video retrieval requests into predefined search categories (Hauptmann et al , 2004, Yan et al ,

2004, Chua et al , 2005) The classification of queries can be based on the general type of visual items requested, such as requests for named persons, named objects, general objects and scenes (Hauptmann et al , 2004, Yan et al , 2004), or can be based on genre specific categories such as person, sports, finance, weather, disaster and general queries for the television news genre (Chua et al , 2005) In both these approaches the topic was identified with the correct query-class through automatic analysis of the query text This approach complements pseudo-relevance feedback mechanisms as it attempts to achieve a better initial ranking of results, whereas pseudo-relevance feedback attempts to improve this initial ranking

The advantage of using query-classes is that it allows a retrieval system to automatically use appropriate retrieval strategies, features and weights for a given topic This enables features that are helpful for specific types of topics to be activated appropriately and both feature selection and fusion weights can be trained using sets of pre-classified topics with associated relevance judgements Alternatively, possibly separate search engines that use different techniques can be used for specific types of queries Without classifying topics, all topics in the video retrieval system are treated the same leading to possible sub-optimal use of the available features

The success of the query classification approach depends on how coherent the video retrieval requests are within their specific categories The most coherent category is finding a specific person, which would most likely make more use of face detection and recognition, speaker identification and specific models for aligning ASR and video OCR named entities with the features The limits are that general categories (general objects and scenes) are extremely broad and these topics are unlikely to be improved by a significant amount from this general classification other than the beneficial effect of having specific categories such as people removed from consideration Yan et al (2004) achieve very good performance for query-class dependent weights on the TRECVid 2003 dataset with a reported MAP of 0 20 which is an 11% improvement on using the single best oracle weight for combining modalities for all topics (MAP 0 18) and considerably better than their text-only (MAP 0 15) and their standard multimodal run (MAP 0 14) They used the same set of topics on a related but separate collection to train the weights of the query-classes and therefore these improvements are likely over-estimated than would be the case if a completely independent set of topics were used to train the query-classes This may explain why the large performance increase was not repeated when they applied the same technique on the TRECVid 2004 search task

3 3 2 Text-Based Video Retrieval

Text retrieval is the most successful retrieval modality for TRECVid video retrieval The ASR text feature is the single best performing feature in successive TRECVid search

experiments

The actual retrieval models for most video retrieval systems are the same as for standard text retrieval and most systems use BM25 (Smith et al , 2002, Hauptmann et al , 2002, Cheng and Chen, 2005), though others use TF-IDF (Heesh et al , 2005) and hierarchical language models (Westerveld, de Vries and van Ballegooij, 2003) Hierarchical language models extend the Jelinek-Mercer (linear interpolation) language model by smoothing the shot model with the adjacent text from a window of shots, the video text, and, as usual for a language model, the background collection text model This model has performed consistently well compared to other models for the TRECVid experiments (Baan et al , 2002, Westerveld, de Vries and van Ballegooij, 2003, Westerveld et al , 2004, Ianeva et al , 2005, Smeaton et al , 2004a)

Smoothing using adjacent text reduces problems caused by the time-delay problem where entities are mentioned before and sometimes after they appear in the related shots Other approaches to handling the time-delay problem include adding the text from adjacent shots to the shot representation (Heesh et al , 2005), combining the retrieval score of the shot with a window of adjacent shots and the video text (Baan et al , 2002), retrieving shots based on speaker segments and mapping back to shots (Quenot et al , 2005), and propagating the shot's score to adjacent shots using a decay function (Chen and Hauptmann, 2004)

3 3 3 High-Level Concept-Based Video Retrieval

Somewhat unlike free text queries, concepts are generally binary in the query when they are used, i e they are either present or not In some systems querying based on the complement of the concept is also supported (Snoek et al , 2005) Query-class approaches to query pre-processing can supplement the query representation with additional concepts that are relevant to the information need such as adding a person and a face concept to a request for a specific person As noted by Snoek et al (2005) concept-based search is very useful when the search topic has a high overlap with an indexed concept Also cross-modal pseudo-relevance feedback can attempt to identify high-level concepts that are correlated with the initial result list and are therefore possibly useful in improving the search results (Hauptmann et al , 2004)

Rautiainen et al (2005) combine 100 semantic concepts from the IBM VideoDIG project [1] in rank-based video retrieval For each concept the top 2000 shots are used and the confidence score is normalised based on rank Queries for multiple concepts sum the normalised confidence scores of the individual concepts Rautiainen et al (2005) found that using 2000 results for each concept under-represents some concepts that occur

[1] Video Dense Information Grding (VideoDIG), http //www research ibm com/VideoDIG/

more frequently and over-represents other features that occur very infrequently This is in fact a general problem with supporting concept-based searching It is necessary to normalise each concept's confidence scores so that they can be reliably combined with each other and with other modalities but to normalise the scores it is necessary to know either the true distribution of the feature within the collection or a reasonable cut-off point for the feature Unfortunately, the cut-off confidence values varies from feature to feature and the frequency of a feature depends on the collection

Concepts can also be used as supplemental information for query-by-example searching Amir et al (2005) combine 46 dimensional semantic confidence weights corresponding to high-level features with low-level features for both query images and shot keyframes Query-by-example is supported by simply calculating the distance between these augmented query and shot vectors

3 3 4 Visual-Based Video Retrieval

While visual features are semantically very different in modelling colour, texture, shape and motion, they predominantly are represented as either vectors or histograms The defacto standard for comparing vector and histogram representations in visual retrieval systems are the Minkowski distance measure and Manhattan and Euclidean distance Many alternatives measures such as Histogram intersection, the information theoretic measures Relative Entropy and Jensen-Shannon distance (Jeffrey divergence), the χ^2-statistic, and the probabilistic measures of query-likelihood and document-likelihood have also been used for computing histogram similarity Measures that take into account the cross-correlation or cross-similarity between histogram bins such as Mahalanobis distance, Quadratic distance and Earth Mover Distance have also been proposed for visual retrieval but are less efficient to compute We will now describe each of these similarity/matching models

Minkowski form distances In visual retrieval, vectors or histograms representing the visual content of the queries and documents are usually compared using Minkowski form metrics such as Euclidean distance (Hauptmann et al , 2004, Pickering et al , 2003, Snoek et al , 2005) and Manhattan distance (Rautiainen, Penttila, Vorobiev, Noponen, Vayrynen, Hosio, Matinmikko, Makela, Peltola, Ojala and Seppanen, 2003) The Minkowski distance,

$$d_{L_p}(\mathbf{q}, \mathbf{d}) = \left(\sum_i |q_i - d_i|^p \right)^{\frac{1}{p}}, \tag{75}$$

is the Manhattan distance (L_1 norm or city block distance) when $p = 1$, the Euclidean distance (L_2 norm) when $p = 2$, and the maximum distance between vector elements (L_∞ norm) when $p = \infty$ Minkowski distances are distance metrics when $p \geq 1$

Puzicha et al (1999) investigated these measures and concluded that overall L_1 was better than L_2 and that both measures were better than L_∞ in their texture and colour classification experiments The superiority of L_1 over L_2 was not universal in all their tested feature configurations and in particular their results showed that for a small LAB colour histogram L_2 marginally out-performed L_1

Fractional distances Fractional distances (Aggarwal et al , 2001) are Minkowski distances with $p \in [0, 1]$ and were recently compared to Manhattan and Euclidean distance for the Corel image classification task, the ImageCLEF image retrieval task and the TRECVid 2003 video shot search task (Howarth and Ruger, 2005) Unlike Minkowski norms, Fractional distances are not a distance metric Howarth and Ruger (2005) find that Fractional distances improve on Manhattan and Euclidean distances for most of the features they tested (RGB histogram, HSV histogram, MPEG7 Colour Structure descriptor and a Convolution feature) Most of the features that perform well for fractional distances are histogram features, while their Gabor and thumbnail features, which are not histogram/distribution features, did not respond as well to fractional distances Howarth and Ruger (2005) explained this response difference between these groups of features as due to the relative density and sparsity of the features – suggesting that dense vectors are less suited for Fractional distances We believe that the difference between features could also be due to the predisposition of distributional features towards lower order Minkowski distances such as Manhattan ($p = 1$) or Fractional distances ($p < 1$) as opposed to Euclidean and Max distance measures Higher order Minkowski norms such as Euclidean or in the worst case the Max norm elevate the effect of large differences in a minority of histograms bins, thereby allowing a minority of histogram bin differences in the distribution to dominate the distance calculation, which may be beneficial for exact matching but is probably undesirable for general visual retrieval tasks Howarth and Ruger (2005) recommend a Fractional distance with $p = 0\ 5$ but find that the optimum value of p varies between 0 25 and 0 75 depending on the feature and test collection

Histogram Intersection The Histogram Intersection between the query histogram H_Q and the document histogram H_D is defined as

$$d_{Int}(H_Q, H_D) = \frac{\sum_i min(H_Q(i), H_D(i))}{\sum_i H_D(i)} \tag{76}$$

Swain and Ballard (1991) proposed the Histogram Intersection measure for comparing colour histograms for image retrieval and similar to query-likelihood only the bins (or discrete symbols) that are non-zero in the query representation contribute to the ranking When histograms are of equal size (or normalised) then Histogram Intersection is equivalent to Manhattan distance (Rubner et al , 2001)

Relative Entropy Relative entropy (or Kullback-Leibler divergence) was suggested in (Ojala et al , 1996) for the image texture classification task We previously defined relative entropy when discussing language models (see Section 2 5 3) Relative entropy measures the average inefficiency of using one distribution to encode another (Cover and Thomas, 1991) It is a directional distance and is more effective for image retrieval when testing the efficiency of encoding the query distribution with the document distribution than the other way around As discussed in the context of language models, relative entropy is undefined when an event to be predicted has zero probability

Jensen-Shannon Distance The Jensen-Shannon distance or Jeffrey divergence/distance (Rao, 1982, Liu, 1991) is a variation of relative entropy (Kullback-Leibler divergence) between two probability distributions and is defined as

$$d_{JS}(P,Q) = \frac{1}{2}D(P||\hat{M}) + \frac{1}{2}D(Q||\hat{M}) \tag{77}$$

where $\hat{M} = \frac{1}{2}(P+Q)$ The Jensen-Shannon distance between two histograms (empirical distributions, i e maximum likelihood probability distributions) H_Q and H_D can be calculated as

$$d_{JS}(H_Q, H_D) = \frac{1}{2}\sum_{i=1}^{n}\left(H_Q(i)\log\left(\frac{H_Q(i)}{m_i}\right) + H_D(i)\log\left(\frac{H_D(i)}{m_i}\right)\right), \tag{78}$$

where $m_i = \frac{H_Q(i)+H_D(i)}{2}$

This particular Jensen-Shannon distances tests the efficiency of assuming that a common source generated both distributions – the query and the document It literally measures the average inefficiency of encoding events from both distributions in the common source Unlike relative entropy it is symmetric and fully defined when comparing two empirical distributions Similar to the language modelling approach to information retrieval, the Jensen-Shannon distance is a measure of the hypothesis that both the query and document are generated from the same source but Jensen-Shannon distance incorporates a measure of how well both the document and the query fit the common source and does not require that the probability distributions be smoothed Recently, it has been established in (Endres and Schmdelm, 2003) that this particular Jensen-Shannon distance is the square of a metric (i e $\sqrt{d_{JS}}$ is a metric), which is significant because it means that it may be possible to use efficient indexing structures to support retrieval (Clarkson, 1997) by using the square root of the Jensen-Shannon distance as the similarity function The square root is of course a monotonic function and therefore ranking in this metric space is the same as ranking with the Jensen-Shannon distance itself Puzicha et al (1997) proposed Jensen-Shannon distance as a dissimilarity measure for unsupervised texture segmentation and texture-based retrieval Puzicha et al (1999) found that this measure generally performs better or at worst similar to L_1 for their colour and texture classification experiments

χ^2-statistic The χ^2-statistic, which was proposed for image retrieval in (Puzicha et al , 1997), can be used to test the hypothesis that both empirical distributions for query and document are generated from the same source The χ^2-statistic is defined for query and document histograms as

$$d_{\chi^2}(H_Q, H_D) = \sum_i \frac{(H_D(i) - m_i)^2}{m_i}, \tag{79}$$

where $m_i = \frac{H_Q(i)+H_D(i)}{2}$ Similar to the specific Jensen-Shannon distance the hypothetical source is modelled as a mid-point distribution Like Euclidean distance, individual differences between histogram bins are squared, which can have the effect of magnifying differences in a few bins so as to dominate the statistic In their colour and texture classification experiments, Puzicha et al (1999) established that this χ^2 distance achieves nearly identical results to the Jensen-Shannon distance

Mahalanobis distance In contrast to previous distance measures, Mahalanobis distance takes into account the statistical correlation between vector/histogram bins The Mahalanobis distance between the query vector V_Q and the document vector V_D is defined as

$$d_{Mah}(V_Q, V_D) = \sqrt{(V_Q - V_D)\Sigma^{-1}(V_Q - V_D)} \tag{80}$$

where Σ is the covariance matrix of the distribution If the covariance matrix is diagonal then this measure simplifies to

$$d_{Mah}(V_Q, V_D) = \sum \frac{(V_Q(i) - V_D(i))^2}{\sigma_i^2}, \tag{81}$$

where σ_i is the standard deviation of the i-th document vector component across the collection In this configuration it is often referred to as normalised Euclidean distance – Euclidean distance itself is achieved when the covariance matrix is the identity matrix The advantage of Mahalanobis distance is that it takes into account the variation in components of the vector when calculating similarity and thereby normalises the effect of different vector components having a different range or variation, which can occur in a severe form when different features are concatenated into a single document vector When all components are independent (diagonal covariance) and have equal variance, this measure produces equivalent scoring as Euclidean distance in terms of ranking documents Chen and Chang (2000) used Mahalanobis distance to retrieve shots based on motion trajectory vectors

Quadratic Distance The Quadratic form distance, which was proposed for comparing colour distributions in (Ioka, 1989) and popularised by the QBIC system (Faloutsos et al , 1994), is defined as

$$D(H_Q, H_D) = \sqrt{(\mathbf{H_Q} - \mathbf{H_D})^T \mathbf{A} (\mathbf{H_Q} - \mathbf{H_D})} \tag{82}$$

where matrix **A** defines cross-bin similarities Similar to Mahalanobis distance, this similarity function is very costly to compute This measure is most often applied to colour histograms, where the entries in the matrix **A** try to capture the perceptual similarities between respective colours Typically the entries a_{ij} in the similarity matrix **A** have been chosen as either

$$a_{ij} = 1 - \frac{d_{ij}}{d_{max}} \tag{83}$$

or,

$$a_{ij} = e^{-\sigma\left(\frac{d_{ij}}{d_{max}}\right)^2} \tag{84}$$

where d_{ij} is the Euclidean distance between the two colour i and j in some colourspace, d_{max} is the maximum of such distances and σ is some positive constant (Hafner et al , 1995) The second equation (Equation 84) increases the fall-off of cross-bin similarity entries for larger distances and larger values of σ make the matrix more diagonally dominant

Earth Mover Distance The Earth Mover distance (Rubner et al , 2000) is based on the minimal cost that must be paid to transform one distribution into the other This measure converts the histogram dissimilarity function into a transportation problem, which is a dynamic programming problem that is costly to compute The benefit of this measure is that it allows local binning of the histograms, which can reduce the storage costs for each image's histogram while retaining relatively the same amount of useful information for retrieval

Document-Likelihood Jin and Hauptmann (2002) present a probabilistic source model (generative discrete document-likelihood) approach to image retrieval in which they assume that an image is generated by a stochastic process They modelled the joint distribution of colour and location as defined by a normalised document-likelihood

$$\begin{aligned} RSV_{Q,D} = \frac{\mathbb{P}(D|Q)}{|D|} &= \frac{1}{|D|} \prod_{(h,v,c,L)\in D} P_Q(h,v,c,L) \\ &= \frac{1}{|D|} \prod_{(h,v,c,L)\in D} \mathbb{P}_Q(h|L)\mathbb{P}_Q(v|h,L)\mathbb{P}_Q(c|v,h,L)\mathbb{P}_Q(L) \\ &\approx \frac{1}{|D|} \prod_{(h,v,c,L)\in D} \mathbb{P}_Q(h|L)\mathbb{P}_Q(v|L)\mathbb{P}_Q(c|h) \end{aligned} \tag{85}$$

where h, v, c are the hue, value and chroma of the Munsell colourspace for each pixel, and L is one of 16 grid locations in the image The h, v, c values are uniformly quantised into 8, 4 and 4 bins respectively based on their distribution throughout the whole collection of images so that each quantisation level contains an equal number of pixels They used two smoothing methods to handle the zero-frequency problem Laplace smoothing (prior of 1 observation for each event) and local smoothing of estimates for each location

using a linear interpolation with neighbouring locations They ranked images by the probability of the query's smoothed probability model generating the document features (document-likelihood), which they normalise by dividing by the document size Though not presented as a language modelling approach, this source probabilistic model has the required qualities such as the estimation of probabilistic models and the use of smoothing techniques The use of independence assumptions between some components of the colourspace (see Equation 85) removes some of the elegance of its realisation and applying document-likelihood with essentially a multinomial distribution probably causes normalisation problems that manifest in terms of the ideal document match as mentioned in the previous chapter (i e the ideal document contains grid regions where all pixels are the colour of maximum probability for the respective query region)

Westerveld et al (2004) present a document-likelihood approach to visual retrieval that models the query images using a Jelinek-Mercer smoothed Gaussian Mixture Model and normalises the document-likelihood with the marginalised probability of the document content within the collection

$$RSV_{Q,D} = \frac{\mathbb{P}(D|Q,r)}{\mathbb{P}(D|\bar{r})} \approx \frac{\mathbb{P}(D|Q)}{\mathbb{P}(D)} = \prod_{x \in D} \frac{\mathbb{P}_Q(x)}{\mathbb{P}(x)} \tag{86}$$

where r and $\bar{r}$ are the relevance and non-relevance events necessarily dropped in the approximation This probabilistic model differs from the probabilistic source model in the normalisation factor of the document-likelihood The results for visual-only searching, showed no improvement for document-likelihood approach with a decrease in MAP reported from 0 028 to 0 026 for full image retrieval and little change for manually selected query image regions with both methods achieving a MAP of 0 026 The incorporation of an automatic method of selecting distinctive samples in the learning of the query's GMM model statistically significantly improved on their query-likelihood method achieving a MAP of 0 034 on TRECVid 2003 (see Table 3)

Query-Likelihood Query-likelihood for a Gaussian Mixture Model (GMM) document representation and approximations of the query-likelihood, such as the random sample likelihood (Vasconcelos and Lippman, 1998) and the asymptotic likelihood approximation (Vasconcelos, 2000), were evaluated for the TRECVid video shot retrieval task (Westerveld, 2004, Baan et al , 2002, Westerveld, de Vries and van Ballegooij, 2003) A smoothed GMM query-likelihood (continuous language modelling approach) was proposed in (Westerveld, de Vries and van Ballegooij, 2003), which combined the documents' Gaussian mixture models with the background model using Jelinek-Mercer smoothing This model can be slow for retrieval due to the number of samples in the query distribution (15 minutes per single image query is quoted in Westerveld et al (2004)) and the asymptotic likelihood approximation quickens the retrieval procedure by approximating the KL divergence between the document and query's Gaussian Mixture

Table 2: Visual and multimodal results for the GMM DCT query-likelihood approach using static image representation and dynamic video content representations for the TRECVid search tasks.

	Text	*Static*			*Dynamic*		
Collection	*-Only*	*Visual*	*+Text*	*%Dif*	*Visual*	*+Text*	*%Dif*
TRECVid 2002	0.1212	0.0287	0.0750	-38.1%	—	—	—
TRECVid 2003	0.1296	0.0281	0.1428	+10.1%	0.031	0.149	+14.9%
TRECVid 2004	0.0680	0.008	0.073	+7.3%	0.010	0.073	+7.3%

Table 3: Visual and multimodal results for the continuous GMM DCT query-likelihood, the discrete multinomial (MNM) DCT query-likelihood and the continuous DCT document-likelihood approach using full image examples, manually selected segments of visual examples and automatically selected distinctive image regions using background model in the GMM training for the TRECVid 2003 search task.

Visual	*Full Examples*		*Manual Segments*		*BG-Train*	
Retrieval Method	*Visual*	*+Text*	*Visual*	*+Text*	*Visual*	*+Text*
Query-Likelihood (MNM)	0.0044	–	0.0066	–	–	–
Query-Likelihood (GMM)	0.0281	0.143	0.0264	0.142	0.018	–
Doc.-Likelihood (GMM)	0.026	0.119	0.026	0.167	0.034	0.162

Model under assumptions that are unfortunately not plausible for the TRECVid collection and which actually degrade retrieval performance compared to query-likelihood (Westerveld, de Vries, Van Ballegooij, de Jong and Hiemstra, 2003). The random sample likelihood also uses GMMs for both query and document feature representations and ranks documents based on the likelihood that a random sample from the query model is generated by the document model. The size of the random sample controls the time required to evaluate this approximation to query-likelihood. Unfortunately, this method is also not as effective as query-likelihood for video shot retrieval (Westerveld, 2004). The MAP performance of the query-likelihood DCT GMM approach for the three most recent TRECVids are shown in Table 2, which represents a baseline performance for these collections that we will compare against in later chapters. More recently, de Vries and Westerveld (2004) compared their continuous GMM approach with a discrete Jelinek-Mercer smoothed language modelling approach for the DCT feature on the TRECVid 2003 collection and found that their discrete smoothed query-likelihood approach was many times worse than their continuous GMM approach (see Table 3).

Rubner et al. (2001) compared the visual similarity measures L_1, L_2, L_∞, χ^2 distance, Jensen-Shannon distance and Relative entropy for colour and texture-based image classification and retrieval. Unfortunately, their image retrieval task is not representative of the general video retrieval task. They randomly selected 94 images from the Corel Stock Photo Library which formed the basis of both their queries and relevant

documents They took 16 disjoint sets of random samples of sizes 4, 8, 16, 32 and 64 pixels from these images and created a ground truth dataset of 1504 image samples for these 94 different query classes A similar procedure was followed with 94 Brodatz texture images to generate their texture queries and relevant document These search tasks, especially the colour experiments, are too artificial and bear little resemblance to retrieving images based on a visual information need or even based on their similarity to complete image examples

Deselaers (2003, Chapter 8) compared the matching models Jensen-Shannon divergence, χ^2 distance, L_1, and L_2 distance for their invariant feature histogram on the WANG (subset of 1000 images from Corel dataset) and IRMA-1617 (subset of 1617 radiographs from Image Retrieval in Medical Applications dataset) datasets and found that L_1 distance, χ^2 distance and Jensen-Shannon divergence achieve similar results and outperformed L_2 distances on both collections

The TRECVid workshop supports the benchmarking of retrieval runs for the video search task but due to the use of different retrieval models, features and fusion methods by participants, it is normally impossible to attribute differences in a submitted retrieval run's performance solely to the choice of the visual similarity/matching model

3 3 5 Fusion

Because video retrieval is situated in a diverse feature environment, it potentially could benefit from the combination somehow of many different features These include text (automatic speech recognition text, closed caption text, video optical character recognition text), audio features (e g monologues, music, gun firing), visual features (colour, texture, shape), motion features (cameras and objects), high-level concepts ('visual keywords' such as outdoors, indoors, landscape, faces) and other specific audiovisual models such as for identifying specific people, animals or objects Early fusion methods, which combine features before performing matching, are not practical for such a large number of features due to the high dimensionality of any combined representation In contrast, late fusion methods perform matching on individual features and fuse these matching scores Late fusion of diverse features is more prudent and can potentially support adaptive fusion methods when relevance information is available

At their most basic, late fusion methods combine the scored and ranked retrieval results from different systems/models/features in order to improve upon the best individual retrieval result Traditional fusion techniques m information retrieval can be broadly divided into rank and score-based (Smeaton, 1998) Rank-based methods such as Borda count combine separate search results based on summing rank positions where the top-ranked of N documents gets a score of N, second gets N-1, etc , and the total score

for a document is the sum of the scores from the different result lists An extension to this combination method is weighted Borda count (linear interpolation of results), which gives preferential weight to specific search result lists Traditional score-based combination methods include CombSUM, which sums the multiple scores, and CombMNZ which sums the scores from truncated result lists (such as top 1000) and multiplies the average by the number of retrieval models that returned it (Lee, 1997) Weights are predominantly included though a linear interpolation of scores When combining heterogenous retrieval models/features normalisation of retrieval scores is necessary and generally involves truncating result lists to the top N results and linear normalising the results from 0 to 1

In the image and video classification task, it is common to use early fusion techniques that combine multiple features into a single vector representation and then optionally apply dimension reduction techniques, while for video retrieval late fusion techniques are the norm and specifically CombSUM, CombWtScore (interpolation of the similarity scores), Borda fuse and round-robin (combine results by choosing the top ranked documents from each result list and repeat this process for subsequent ranks removing duplicates) are the most common Due to research groups using different features and retrieval models it is difficult to know which fusion strategy works best overall simply from their submissions to TRECVid Likewise, it is difficult to compare different feature sets and retrieval models when the fusion approaches are dissimilar It is useful to think of fusion as occurring at many separate levels in video retrieval systems such as for combining text features, combining visual features, combining multiple query examples, combining multiple document representations (keyframes/keysegments), and combining multimodal features Different fusion strategies may be best or more stable for different fusion tasks

Combining text features Adams et al (2003) combined multiple text retrieval scores by summing normalised text scores Most fusion strategies have originally been applied to the task of combining results of multiple text search engines and can directly be applied to this task

Combining visual features Visual features can be combined using early fusion or late fusion techniques Rautiainen, Penttila, Vorobiev, Noponen, Vayrynen, Hosio, Matinmikko, Makela, Peltola, Ojala and Seppanen (2003) combine their different feature vectors before matching by appending the vector representation of each feature vector Hauptmann et al (2002) combine regional colour and texture features into a single vector representation and then reduce it to 50 dimensions by using singular value decomposition Early fusion techniques are less popular for the video retrieval task due to the potentially high dimensional representations and the need to support dynamic

reweighting of features due to explicit direction by the user in the user interface or due to relevance feedback Hauptmann et al (2003) combine their colour representation with their texture representation using a weighted interpolation of probabilities The probabilities were derived from ranks so this method resembles a weighted Borda fuse Rautiainen, Ojala and Seppanen (2004) combined visual feature by summing the scores, while Pickering et al (2003) combine visual dissimilarity scores for different feature vectors using a linear interpolation of scores Quenot et al (2005) also use a linear interpolation of scores to combine results but allow the users to adjust the weights in the user interface

Combining multiple query examples Different fusion approaches have been applied to the task of fusing multiple query examples within the same topic such as CombSUM (Hauptmann et al , 2002), Round-Robin (Westerveld et al , 2004) and independent probability (Westerveld, de Vries and van Ballegooij, 2003) Adams et al (2003) support the selection of the following fusion methods for combining the scores of multiple visual examples average scores (or equivalently sum score), minimum score (CombMIN), maximum score (CombMAX) and product of scores

Combining multiple keyframes The retrieval results of multiple keyframes for the same shot can be fused with the same methods that are used to combine multiple query examples The most popular method for combining the scores of multiple keyframes from a single shot is to take the maximum of the keyframe scores (Hauptmann et al , 2002, Quenot et al , 2005)

Combining multiple modes Hauptmann et al (2002) combine face detection, speaker identification, OCR, ASR and image retrieval similarity scores using a linear interpolation of scores Heesh et al (2005) also combined their text and visual features using a weighted linear interpolation of individual scores but in their manual experiments they found no benefit to visual retrieval when combined with text search – in fact it decreased their retrieval performance compared to text alone Baan et al (2002) combine their text language model results with their DCT GMM visual model using independent probability assumption (multiply probabilities or sum log-likelihoods) Weighted Borda fuse of the multimodal results is used for multimodal fusion in (Rautiainen, Ojala and Seppanen, 2004, Hauptmann et al , 2004) Another method of combining multimodal results is to use one of the features as a filter on the results of the other Rautiainen, Penttila, Vorobiev, Noponen, Vayrynen, Hosio, Matinmikko, Makela, Peltola, Ojala and Seppanen (2003) use the text and concept results to filter the visual results Yan and Hauptmann (2004) use a boosted co-training approach that trains the weights for combining concept and low-level feature results with text-based results on a per-query basis

Strongly related to the multimedia fusion task is the use of query-classes to automatically select different features and fusion weights depending on the type of query (Yan et al , 2004)

In a theoretical study on fusion strategies in video retrieval, Yan and Hauptmann (2003) establish upperbounds for the performance of fusion models They consider upper bounds for monotonic combination functions and for the more restrictive linear combination functions There is a lack of empirical studies of the different fusion methods for the different video retrieval fusion tasks that we have identified in this section Yavlinsky et al (2004) compared CombMIN, CombMAX, CombSUM and Borda fusion for the task of combining text and their visual features' results on TRECVid 2003 but found that no fusion method improved on the results of text alone The Lowlands group at TRECVid found that their fusion strategy for combining their hierarchical text language model and the GMM DCT visual feature improved on text-only results for TRECVid 2003 and 2004 but failed on TRECVid 2002 (see Table 2 in previous section) Most of the submitted runs in the TRECVid experiments are either from interactive retrieval experiments or from manual experiments that permit the modification of the topic representation and therefore cannot be compared across retrieval groups to elucidate the relative benefits of different fusion strategies

3 3 6 Relevance Feedback

Relevance feedback has gained much attention in recent years in the image retrieval community We will briefly describe a select few image and video relevance feedback approaches — a more comprehensive review is available in (Zhou and Huang, 2003) MARS (Rui, Huang and Mehrotra 1997) and MindReader (Ishikawa et al , 1998) are two of the baseline implementations of relevance feedback in image retrieval In each relevance feedback iteration both of these systems move the query model closer to the centre of the cluster of relevant documents and modify the importance of different components of the feature representation in the calculation of similarity between documents and the updated query model

MARS is one of the first image retrieval system to support relevance feedback and incorporates two feedback strategies (Rui, Huang and Mehrotra, 1997) Firstly, similar to Rocchio the original query vector is updated so that it is closer to the relevant documents and further away from the non-relevant set of documents Secondly, it re-weights different elements of the document vector representation depending on its variance within the relevant set of documents Specifically it weights a vector component by the reciprocal of its standard deviation

MindReader learns the pertinent features and their relationships for a set of relevant examples (Ishikawa et al , 1998) Both MARS and MindReader essentially use a quadratic distance measure between the query model and the document representations The difference between the two systems is that MARS only uses the diagonal elements of the quadratic distance matrix, whereas MmdReader uses the full quadratic distance matrix MindReader requires a lot of relevance judgements to properly estimate this quadratic distance matrix and enough relevance judgements are unlikely to be available for large feature representations

More recently, Giacinto and Roli (2004) suggest that the global representation of relevant documents in terms of a single unimodal query model suffers due to the small number of relevance judgements and the possibly disjoint nature of relevant documents within the feature space They suggest using the ratio of minimum distance from relevant documents to non-relevant documents in updating the ranking for each feedback iteration This nearest neighbour approach is different to MARS or MindReader, which try to represent the query as a single query-modal vector in the feature space

The TRECVid topics' image and video examples provide a means for using relevance feedback techniques even before a user has started to indicate which shots from the collection are relevant For example, it may be possible to use a topic's initial visual examples to select features and weight their importance Within the submissions for the TRECVid search task, a few systems have integrated relevance feedback

In TRECVid 2002, the IBM search system (Adams et al , 2003) performs relevance feedback by partitioning the feature space with a piecewise linear decision surface that separates the relevant and non-relevant shots where each of the piecewise decision surfaces are normal to the minimum distance vector from a non-relevant point to the convex hull of the relevant points (Ashwin et al , 2002) This feedback mechanism showed some promise for the image retrieval task in a relatively low-dimensional feature space (Ashwm et al , 2002) but has not been evaluated on the TRECVid search task

Yan et al (2003) proposed a negative pseudo-relevance feedback approach for content-based video retrieval The positive examples were the original query examples and the negative examples were the low ranking results from the initial search They used a Support Vector Machine (SVM) to generate the feedback scores for the documents, which were combined with their initial result scores to produce the updated retrieval results The feature space for feedback consisted of HSV colour and Gabor texture information Their evaluation of this method on the TRECVid 2002 search tasks showed positive improvement (MAP 0 1522, 7 6% improvement) compared to their baseline non-feedback system In total, 14 of the queries were improved and only 4 disimproved in terms of average precision on the 25 TRECVid 2002 topics (Yan et al , 2003)

Related to pseudo-relevance feedback is Co-Retrieval (Yan and Hauptmann, 2004),

which automatically selects and weights additional weak features based on an initial ranking of results Co-Retrieval is a modification of co-training (Blum and Mitchell, 1998) to take into account unlabelled documents and is based on the premise that an independent and redundant feature split improves the performance of co-training (Nigam and Ghani, 2000) They split their features into two sets the first set is typically text features and the second set are low level features (colour, edge and texture histograms) and semantic features such as from face, anchor, commercial and outdoor feature detectors Their Co-Retrieval approach first uses the first set features (i e more reliable text features) to create a set of pseudo-relevant document and then uses these results to select and train the weights of the second set of weaker features They learn the weights of the second set by using a boosting type algorithm and a noisy label prediction scheme Co-Retrieval was evaluated on the TRECVid 2003 search task where it achieved a MAP of 0 208, 10% better than using the global optimum weight across the feature set for all topics

3.4 Evaluation of Video Retrieval using TRECVid

The Text REtrieval Conference (TREC) is an annual conference organised by the U S National Institute of Standards and Technology (NIST) to bring together both academic and commercial researchers in the field of information retrieval and is separated into tracks each focussed on a specific IR discipline Originally the TREC conference was solely focussed on text retrieval but it has expanded to include amongst others, web, spoken document, multilingual and more recently video retrieval

The TREC Video Track (TRECVid) was first held in 2001 as part of the TREC conference and has recently completed its fourth annual cycle of operation The popularity and importance of TRECVid has grown over the years - originally in TRECVid 2001 there were 12 groups participating and more recently in TRECVid 2004 this has grown to 33 The evaluated tasks have also increased from shot boundary detection and interactive and manual search in TRECVid 2001 to also include feature detection (TRECVid 2002 onwards), news story segmentation (TRECVid 2003 and TRECVid 2004) and a pilot of fully automatic search (facilitated in TRECVid 2004 but first supported in TRECVid 2001) The shot boundary task concerns the correct identification of different types of shot boundaries such as cuts, fades, dissolves, while the feature extraction task concerns the identification of general concepts that might be useful for information retrieval and finally the news story segmentation task concerns identifying the semantic news stories within broadcast television news Since in this thesis we are primarily interested in video retrieval, we will describe the activities of TRECVid in terms of its video search experiments More information on each year's TRECVid results and the other tasks is available in (Smeaton et al , 2002, Smeaton and Over, 2003, Smeaton et al , 2004b, Kraaij et al , 2004, Smeaton et al , 2004a)

The TRECVid evaluation procedures for video search are grounded within the traditions of controlled experiments in information retrieval that has its roots in the Cranfield experiments (Cleverdon, 1967) and the main TREC initiative Each year research groups evaluate their video retrieval approach (retrieval system, retrieval model, feature representation, user interface, etc) using a supplied reference collection and common set of topics (queries) and submit their results for evaluation by independent NIST assessors These assessors pool the results from multiple groups, make relevance judgements and distribute the results in time for the TRECVid workshop

The first TRECVid experiments, TRECVid 2001, was a dry run for applying the TREC model to evaluating video retrieval The main problems with TRECVid 2001 were that it consisted of only 11 hours of content compared to over 40 hours in subsequent years, the topics were originally suggested by the participants and therefore possibly biased towards the proposers' group and the evaluation measures were not standard information retrieval measures but were based overlap of video clips with the relevant segments The success of TRECVid 2001 was in focusing a community of researchers into a common empirical evaluation framework for video retrieval and also in the improvements in the experimental setup in subsequent years Due to the noted problems with TRECVid 2001 we concentrate our discussion on the larger collections and better experimental setup employed in TRECVid 2002, 2003 and 2004

In the following subsections, we will present the major components of the TRECVid laboratory-based retrieval experiments namely the reference collections, search topics, relevance judgements, experimental setup and evaluation measures

3 4 1 Reference Collection

The TRECVid reference collection is the set of documents to be searched and consists of the videos themselves, the shot definitions and automatic speech recognition (ASR) text Keyframes for each shot were also supplied but it is not mandatory to use them Some research groups donated results for high-level feature detection that could be used by other groups doing the search task The retrieval unit for the experiments is the video shot and this necessitates a common shot boundary to be distributed with the video collection

The TRECVid 2002 collection consists of 176 advertising, educational, industrial and amateur videos from 1915 to 1975 sourced from the Internet Archive[2] and the Open Video Project[3] (Smeaton and Over, 2003) The visual quality of many of these videos is poor and the overall quality of the videos in this collection is highly variable

[2]Internet Archive, URL http //www archive org/movies/
[3]The Open Video Project, URL http //www open-video org/

The dataset was split into 40 hours for the video search test collection, 5 hours for feature test collection and 5 hours for shot boundary test collection The retrieval unit for the search task is video shots from a common shot boundary definition supplied by CLIPS-IMAG

The TRECVid 2003 and TRECVid 2004 collections contains recent television news programmes which were broadcast from 1998 to 2001 on the ABC, CNN and C-SPAN channels The visual and audio quality of the video is consistent and is of a far higher standard than the previous years collection The search test collection for TRECVid 2003, the subset of the collection for evaluating the video search task, contains 113 news programmes in which 53 are ABC World News Tonight from 1998, 54 are CNN Headline News also from 1998 and 6 are C-SPAN news from 2001 The TRECVid 2004 search test consists of 70 hours of ABC, CNN and C-SPAN from 1997 (Kraaij et al , 2004) ABC and CNN news follow a similar evening news format whereas the C-SPAN news is televised debates of the US Congress A common shot boundary supplied by CLIPS-IMAG and ASR text provided by LIMSI (Gauvain et al , 2002b) was used by participants in the video search task experiments

The purpose of the reference collection is to recreate a real-world situation in a static collection that is large enough so that results based on experiments on the reference collection can be expected to apply to the real-world scenario As mentioned previously, the first search test collection in TRECVid 2001 failed many of these criteria but served as a dry run of the TRECVid evaluation framework TRECVid 2002 provided a better sized collection but because of its visually varied nature (i e poor visual encoding) it is difficult to work with The more recent TRECVid 2003 and TRECVid 2004 television news collections are of a similar size but are more representative of current video search applications in terms of quality and purpose While very small compared to operational news collections they provide a challenging retrieval scenario and in terms of content is similar in nature to real-world news collections

3 4 2 Search Topics

Many early image retrieval experiments used search topics (patches of texture or colour images) that had very limited application to general ad hoc retrieval scenarios The TRECVid search topics are motivated from requirements of a hypothetical professional user searching a video collection They are not composed by real end-users of a video retrieval system but are based on the types of queries found in studies of professional visual retrieval environments (Armitage and Enser, 1996), which classified visual retrieval requests into amongst others general/specific requests for person, things, events or places We therefore cannot claim that a percentage improvement in a TRECVid search experiment is transferable directly to a specific set of real users, but significant improvements

Figure 17: The first TRECVid 2002 search topic with all relevant shots displayed.

in TRECVid should be indicative of a potential improvement in some real-world video retrieval application and more extensive field testing in specific domains with real users would be required to completely validate claims.

Each TRECVid topic consists of a short text description with optional image, video and/or audio examples. In TRECVid 2002 and onwards NIST composed the queries, which contain for example requests for the following content:

- 103. Find shots of Yasser Arafat (specific person)
- 110. Find shots of a person diving into some water (generic person, generic event/action)
- 106. Find shots of the Tomb of the Unknown Soldier at Arlington National Cemetery. (specific thing, specific place)
- 130. Find shots of a hockey rink with at least one of the nets from some point of view (generic thing, generic place)
- 139. Find shots of a handheld weapon firing (generic thing, generic event/action)

The full topic list for TRECVid 2002, 2003 and 2004 are available in Tables 28, 29 and 30 in the appendix (pp. 232, 233 and 234). The full topic description including text, image examples and video examples are shown Figures 17 and 18 for the first topic in TRECVid 2002 and TRECVid 2003. It should be recognised from these two examples that the visual examples for a particular topic may not be visually coherent. We can see this if we look at Figure 18, where the keyframes of relevant shots do not necessarily contain the relevant items, though the video shot it represent does.

Figure 18: The first TRECVid 2003 search topic with the first 16 of 87 relevant shots displayed.

Table 4: Distribution of topics within each topic classification type for the TRECVid search tasks.

Search Col.	*Specific/Named*			*Generic*		
	Person/Thing	Event	Place	Person/Thing	Event	Place
TRECVid 2002	8 (32%)	0 (0%)	4 (16%)	18 (72%)	10 (40%)	5 (20%)
TRECVid 2003	9 (36%)	0 (0%)	1 (4%)	16 (64%)	9 (36%)	4 (16%)
TRECVid 2004	7 (30%)	0 (0%)	0 (0%)	17 (74%)	12 (52%)	3 (13%)

In Table 4 we summarise the categorisation breakdown of topics for the three TRECVid collections, which is also indicated in the tables in the appendix for each topic The number of topics in each category is consistent for all but the specific place category, which is only represented by one topic in TRECVid 2003 and none in TRECVid 2004 The number of topics involving event/action was increased at the request of participants in TRECVid 2004 About a third of topics are for specific people or things, the other two thirds concern generic things or people and between 40% and 50% of topics also specify a generic action/event Nearly all recent topics from TRECVid 2003 onwards do not mention a specific place and only 13% to 20% of topics mention a generic place in their topic description The proportion of TRECVid topics in the different categories does not match the distribution in real-world video collections such as the BBC Natural History Unit or the British Film Institutes's Natural Film and Television Archive, which contain about double the proportion of requests for specific people and things and at least half the proportion of requests for generic person and things (Smeaton and Over, 2003) However, the topics are quite varied and somewhat represent the general types of queries found in real world applications

The high level goal in selecting topics is that they fit within the type of requests of a professional user and roughly correspond to the types of topics in real-world video libraries Other criteria such as ensuring that there are multiple relevant shots in preferably multiple videos and that the topic should not be too difficult are also used in selecting and composing the topics Video search is already very difficult with the currently selected TRECVid topics never mind with more elaborate requests The procedure for selecting topics for TRECVid 2003 and TRECVid 2004 is as follows (Kraaij et al , 2004) Videos in the reference search test collection are viewed without the audio to identify candidate topics The composed topic text therefore is not affected by the dialog in the video The visual examples were also chosen without reference to the relevant shots and when there were more candidate visual examples than required the final set was chosen randomly The procedure for TRECVid 2002 was slightly different in that the audio was not turned off when identifying candidate topics from viewing the collection and therefore it is believed that the results are biased towards ASR-based retrieval since some of the words from the audio dialog were incorporated into the topic description (Smeaton et al , 2002) Also in TRECVid 2002 the visual examples were chosen to be somewhat similar to the preliminary relevant shots so as to try to alleviate some of the difficulty in visual search but as a side effect this potentially biases the results in the visual domain Both these issues were resolved in later TRECVids

Table 5 summarises the statistics of the topics in terms of number of image examples, video examples and relevant shots for the three TRECVid search tasks The average number of visual examples per topic at 5 2 and 6 1 in TRECVid 2003 and TRECVid 2004 is nearly double the amount for TRECVid 2002 Each topic in TRECVid 2004 has a minimum of three visual examples compared to a minimum of one for the other

Table 5: Summary statistics for TRECVid topics showing the average, variance (in brackets), minimum and maximum (in square brackets) of the number of visual examples, image examples, video examples and relevant shots per topic in the TRECVid 2002, 2003 and 2004 collections.

Search Col.	*Vis. Examples*	*Images*	*Videos*	*Relevant*
TRECVid 2002	3.0 (1.3) [1 - 5]	0.6 (1.2) [0 - 5]	2.3 (1.5) [0 - 5]	58 (68) [3 - 303]
TRECVid 2003	5.2 (2.1) [1 - 9]	2.8 (1.3) [0 - 5]	2.3 (1.8) [0 - 6]	85 (133) [6 - 665]
TRECVid 2004	6.1 (2.1) [3 - 11]	2.4 (1.7) [0 - 5]	3.7 (1.3) [1 - 6]	78 (47) [16 - 194]

two collections. The number of video examples has also increased to on average 3.7 videos per topic in TRECVid 2004 up from 2.3 in the other collections. The number of relevant results per topic is more consistent in TRECVid 2004 having a smaller range with between 16 and 194 topics relevant per topic. The TRECVid 2002 search task contains a topic (77. Find pictures of George Washington) with only 3 relevant results and this can cause problems in the evaluation measures as it can achieve very high recall (or MAP) relative to the other topics by the occurrence of any of its relevant shots within the top ranking results of a tested retrieval run. TRECVid 2003 has a topic (117. Find shots of one or more groups of people, in a crowd, walking in an urban environment) with 665 relevant shots and this can also cause problems in the evaluation measures since this topic can unduly influence mean precision at low document cut-offs for this collection.

3.4.3 Relevance Judgements

The relevance judgements are the set of relevant shots for each topic. Relevance judgements are made for the topics after each group has submitted their retrieval runs to NIST. The top X results are pooled from each submitted run and evaluated by relevance assessors. The relevance accessors review the candidate shot based on the topic description and try to make an objective decision on whether the shot is relevant or not. The relevance judgement is a binary decision and simply depends on the information need being present and recognisable within the shot. This decision is not made based on the keyframe but on viewing the full shot.

In TRECVid 2002 all runs were evaluated to a depth of $X = 50$. Pooling in TRECVid 2004 was slightly different to previous years which evaluated using fixed depth pools. In TRECVid 2004 variable depth pools were employed and NIST accessors continued evaluating a topic's pooled results until they ran out of time or stopped finding relevant results. Each topic was evaluated to a minimum depth of 50 results per pooled run. The effect of pooling and judging depth has not been fully evaluated (Kraaij et al., 2004).

This pooling strategy requires a reasonable level of performance for the topics but

many of the topics have very poor performance for manual and fully automatic searching and therefore interactive user experiments play an important secondary role in identifying relevant shots that otherwise may not be detected

3 4 4 Search Experiments

There are three search experiments supported in the TRECVid evaluation fully interactive search, manual search and fully automatic search The experiments differ in the role users play in the process

Interactive search The interactive search task simulates the case where a user is given a fixed time limit - a maximum of 15 minutes in the case of official TRECVid experiments - to find as many relevant shots to the information need as he or she can, using whatever tools and in whatever ways deemed useful This means that the user brings in real-world knowledge and experience The purpose of interactive search experiments is to support user-centred experiments in a controlled environment for comparing for example different interfaces, retrieval strategies or underlying retrieval systems When experiments involve multiple users it is useful to employ a block design so as to reduce the bias It is good practice to supplement performance information such as precision and recall with questionnaires to gather information on the users' experiences with the different test systems and topics

Manual search The manual search task allows a professional user to translate the TRECVid topics into a form suitable for a video retrieval system They can embed into the search request all the subtleties of the topic that the retrieval system supports The purpose of manual search is to allow groups to compare variants of their video retrieval system While the addition of the user in the process makes it possible to better take advantage of the functionality in each system variant, it also adds a topic translation process that is a source of noise and hampers cross-group comparisons

Fully automatic search In the fully automatic search task the retrieval system takes the TRECVid topic without modification by a user The system may then modify it, such as performing stopword removal and stemming for the text description, but all modifications of the topic are performed automatically without user intervention A trial run of the fully automatic search was facilitated in the TRECVid 2004 cycle, after it was abandoned after TRECVid 2001 due to its perceived level of difficulty being too high Fully automatic experiments, which potentially decrease the reported performance compared to manual TRECVid experiments, provide a more unbiased view of system performance that can be compared more easily and repeated independently within the

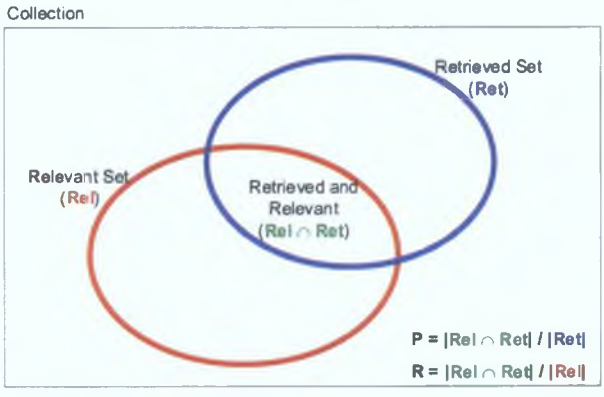


Figure 19: Precision (P) and Recall (R).

research community. The purpose of fully automatic search is to support system variant evaluation in a controlled, repeatable and as bias-free as possible experimental setup. This is the most difficult of the three search experiments in terms of performance measures as there is no human involved in helping the system out. Unlike the other two search experiments this allows for retrieval models or features to be compared across retrieval groups and allows performance to be measured and compared between groups without having to worry about the differences in users. In TRECVid 2004 the fully automatic runs were evaluated as supplemental runs and classified as fully automatic in order to distinguish them. Previous to this, automatic runs that were submitted to TRECVid 2002 and 2003 were classified as manual runs and therefore unfairly compared.

3.4.5 Evaluation Measures

In this section we present the evaluation measures that are used to compare TRECVid runs, which consist of a ranked list of shots for each topic. The evaluation measures quantify the effectiveness of the retrieval system for finding relevant content and are usually calculated at two levels: for each individual topic and an average over all topics in the submitted run.

Recall and Precision are standard information retrieval measures that form the basis of most measures of effectiveness in video retrieval. Recall and Precision are set-based evaluation measures (see Figure 19) that do not take into account the ranking of the search results and are defined for a specific topic as:

Precision Precision is the proportion of retrieved documents that are relevant.

$$Precision = \frac{|Rel \bigcap Ret|}{|Ret|} \tag{87}$$

Recall Recall is the proportion of the total relevant documents that are retrieved

$$Recall = \frac{|Rel \bigcap Ret|}{|Rel|} \tag{88}$$

Both these measures can be averaged over a set of topics in order to achieve the overall precision and recall for a retrieval run Precision relates to a user's preference for results with relevant documents over non-relevant documents, while recall relates to a user's preference for an exhaustive set of relevant documents For result sets, precision and recall have to be reported together in order to get a proper interpretation since a result set with a single result that happens to be relevant has a precision of 1 0 and a result set containing all documents has a recall of 1 0 The effectiveness of a retrieval system is hard to characterise and compare when expressed as a recall and precision pair

Precision is easier to assess correctly than recall and the effort involved is related to the result set length Recall is more difficult to calculate than precision as it requires knowledge of all relevant documents within the test collection and for large collections it is prohibitive expensive to evaluate the relevance of each document As mentioned previously, the strategy employed in TRECVid (and generally in TREC and other TREC-like community-based IR evaluation initiatives) is to pool results from multiple groups, evaluating them to a certain depth in order to achieve an adequate estimate of all relevant documents

It is important to keep in mind that the choice of evaluation measure should be correlated with the potential user's perception of effectiveness The precision measure tells us how correct the results are and is desirable from a users perspective as they would prefer not to wade through many incorrect results in order to find relevant ones High precision is beneficial for most searches but there are occasions when users would tradeoff precision in favour of high recall such as when compiling or researching for a historical documentary

Precision at documents cut-offs This measure provides the precision for subsets of top retrieved documents In TRECVid this precision measure is reported for the top 5, 10, 15, 20, 30, 100, 200, 500 and 1000 documents It is a useful evaluation measure as precision at cut-offs between 10 and 30 correlate with the quality of results that the users will see in the first page of their video search results

Average Precision Average Precision is a single measure of the effectiveness of a retrieval run for a topic's ranked results and is the standard method for comparing topic results in TRECVid It is a rank-sensitive measure that rewards runs that locate relevant results higher in the rankings Average Precision is calculated as the average of

precision when each relevant documents is found in the ranked list with non-retrieved relevant documents giving a precision of zero Average Precision never decreases as more documents are added to the bottom of an existing rank

Mean Average Precision Mean Average Precision (MAP) is used in TRECVid as a single measure for the effectiveness of a search system over a set of topics It is calculated by simply averaging the Average Precision values for each topic

Interpolated precision Interpolated precision is generally used to calculate precision at 11 recall levels (0, 0 1, 0 2, , 0 9, 1 0) with the rule that the interpolated precision at a recall level is the maximum precision at the recall point greater than or equal to the level For example a topic with 3 results would have recall points corresponding to 0 33, 0 67 and 1 0 recall The interpolation precision at the fixed 11 recall levels are *interpolated* from the topics' actual recall points with the previously mentioned rule As for other measures, interpolated precision can be averaged over the full set of topics to achieve a performance measures of the complete retrieval run The standard 11 interpolated precision points are used to plot precision-recall graphs (interpolated precision on Y axis, recall level on X axis) in order to support the graphical comparison of retrieval runs The ranges of these graphs between 0 to 0 2, 0 2 to 0 8 and 0 8 to 1 0 characterise the performance of retrieval runs at high precision, middle recall and high recall (*Common Evaluation Measures*, 2003)

The following statistical tests are not used by the standard TRECVid evaluation but can be employed to test whether differences between retrieval runs in terms of their MAP or other measures are statistically significant These statistical tests can only be calculated over a set of topics not a single topic and both retrieval runs being compared must have results for the same set of topics If a result is statistically significantly better than another then this implies a consistent improvement in the retrieval results that is unlikely to be the result of chance Each significance test, based on particular assumptions, calculates the probability of chance (p-value) accounting for the difference between retrieval runs If this p-value is less than a preset significance level (normally 0 05) then it is assumed that the differences between retrieval runs are statistically significant A range of statistical tests for comparing two or more retrieval models is described in (Hull, 1993), the most common in use in information retrieval are the Sign test and paired Wilcoxon test Detail of their calculation and indeed other statistical tests can be found in (Hull, 1993)

Sign test The paired Sign test only takes into account the sign of the difference between the paired retrieval runs' topic results In each pair of topic results, the difference

between retrieval runs is calculated and a count is made of the number positive differences The decision to reject the null hypothesis (i e assume statistical significant differences) is based solely on the count of positive differences

Paired Wilcoxon test (Wilcoxon signed-rank test) The paired Wilcoxon test takes into account the magnitude of the difference between retrieval runs as well as the sign of the difference It assumes that differences for all paired topic results are distributed from a symmetric continuous distribution This assumption is not strictly valid in information retrieval and leads some researches to advocate the less powerful Sign test (van Rijsbergen, 1979, chap 7) The procedure for calculating the test statistic is to replace each paired difference with the rank of their absolute difference and to multiply this rank by the sign of the difference The sum of these positive and negative differences determine whether the retrieval method is actually statistically significantly better

3.5 Summary

Video retrieval is difficult due to the lack of authoritative semantic features that describe what is perceivable within video In video retrieval, visual features are detached from the semantics of the visual content and even transcribed spoken words (ASR or closed captions) are unreliable due to the time-delay problem and the fact that what is spoken about is not necessarily shown in the accompanying video stream

Video retrieval is complicated due to the large number of indexing processes that must be performed in order to support retrieval The two key indexing processes are video segmentation and feature extraction Video segmentation at a minimum consists of shot segmentation and keyframe extraction but can also include story segmentation Unfortunately, semantic object segmentation and extraction is beyond the ability of current automatic segmentation algorithms for general video content The extracted visual features provide support for query-by-example searching and can be classified into colour, texture, shape, faces, motion and spatio-temporal features High-level features such as people, beaches, city, landscape can be extracted to provide a more semantic and directly meaningful description of the content, though their usefulness depends on how accurately they can be identified and how well they align with the needs of video searchers Extracted speech from the audio stream provides a strong semantic description but other high-level audio features such as music, monologues and dialogue may be of some use

The low-level visual features such as colour, edge, texture, motion, faces and spatio-temporal features can be extracted and represented in may different ways Many of the

feature representations are basically discrete probability distributions of certain events (histograms, correlograms, coherence matrix) and are therefore amenable to interpretation within the language modelling approach to information retrieval as described in the previous chapter The MPEG7 standard also describes many visual features and represents the accumulated research into best practice for efficient and compact features for the video retrieval task Many of the MPEG7 visual features are also interpretable as discrete probability distributions and therefore our advocated language modelling approach will also apply to a range of standard features in MPEG7-based video retrieval systems

We have studied the different types of retrieval models that are used to compare features Specifically we looked at similarity/matching models for histogram and vector representations Though we have identified a quite varied set of similarity models, in practice the overwhelming majority of current video retrieval systems are based on simple geometric distances such as Euclidean distance or Manhattan distance for comparing visual feature vectors Quadratic distance is also popular in image retrieval, though due to the fact that it is relatively expensive to evaluate it is not used as much Statistical tests such as χ^2-distance, Jensen-Shannon distance and probability methods such as document-likelihood and query-likelihood have received much less attention Query-likehhood for the continuous GMM DCT feature representation is the exception, which has been well studied on the TRECVid collections and whose results will form a baseline for comparison in future chapters The GMM approach to modelling video features is more popularly used in video classification and high-level feature detection than in video retrieval due to the time taken to evaluate the continuous GMM query-likelihood

Fusion of retrieval results from different feature representations plays a significant role in determining the effectiveness of a video retrieval system Existing video retrieval systems use a multitude of different fusion approaches There are many fusion tasks such as combining features in single visual example searching, combining multiple visual examples and combining multimodal features, but it is unclear as to which of the fusion techniques is actually best for these tasks Some studies have been performed but have produced negative results for the video search task in terms of combining text and visual features

Relevance feedback has had significant interest in the image retrieval community and though we have not described all current video retrieval relevance feedback research, there is surprisingly little This will likely change in the near future especially with the availability of reference video collections, topics and relevance judgements

Benchmarking of components of video retrieval systems can progress video retrieval research and is supported by the recent TRECVid initiative In the past video test collections and topics were extremely limited both in size and the quality of replicating

a real-world video retrieval task Many studies of visual retrieval before TRECVid collections, were based on retrieving similarly textured or coloured image tiles from image categories Even recent image retrieval research is often based on image categories from the Corel datasets that more mimic a categorisation of content into homogeneous classes than an attempt to meet a specific information/visual need The homogeneity of such classification-based 'retrieval' classes in some particular feature space may result in techniques being advanced that cannot handle the heterogeneity in feature spaces for the more general video retrieval task

The strengthening of video retrieval research in terms of controlled experiment-based investigations is facilitated by the TRECVid initiative, which provides the infrastructure for empirical analysis of video retrieval approaches and systems TRECVid is not perfect and indeed results from it can be misleading in some cases Caution must be exercised when extrapolating results from a single TRECVid collection as these typically contain only 25 topics Therefore researchers should consider using multiple TRECVid collections in their experiments, which is only possible to do recently Tuning of free parameters is an issue and arbitrary choices of parameters may hinder or benefit one method over another Due to TRECVid participants justifiably using different features, retrieval models, fusion approaches and to a lesser extent relevance feedback, interpretation of differences in official TRECVid submissions between research groups is extremely difficult except at a basic level of whether their system configuration performed better or worse than another

CHAPTER IV

PROPOSED DISCRETE LANGUAGE MODELLING APPROACH FOR VIDEO RETRIEVAL

In this chapter we propose a discrete language modelling approach for visual-based video shot retrieval We extend the current text-based language modelling approach to video shot retrieval by considering hierarchical versions of the Dirichlet, Absolute and Witten-Bell language models for the physical and semantic hierarchical structure of video We extend the application of the text-based language modelling approaches to visual-based retrieval by converting the visual features, such as colour, edge and texture, into discrete representations suitable for the originally text-based language modelling approach We propose to combine results for the text and visual features using data-fusion methods originally developed for combining the results of multiple text search engines Additionally in this chapter, we outline our evaluation methodology for the proposed retrieval models, visual features and fusion methods

4 1 Introduction

4 2 Proposed Text Language Model Extensions

4 3 Proposed Visual Language Models

4 4 Fusion Methods

4 5 Evaluation Methodology

4 6 Related Research

4 7 Summary

4 1 Introduction

In this thesis, we deal with the research problem of ad hoc video shot retrieval using a multimedia query consisting of text and multiple image and video examples We evaluate our approach using fully automatic experiments on the TRECVid 2002, TRECVid 2003 and TRECVid 2004 search tasks The TRECVid search topics were chosen by NIST to mimic the type of information needs that occur in a real world video retrieval scenario and concern people, things, events and locations

We propose to use discrete language models for both the text and visual retrieval of video shots We convert features from both mediums into the same representation – a bag of discrete symbols For text the discrete symbols are the document's stemmed words after the stopwords have been filtered out, while for visual features the discrete symbols are the multidimensional histogram entries of the feature's discrete representation We convert a continuous visual feature into a discrete representation by simply quantising each of its dimension's values and by truncating the number of dimensions

for high-dimensional features

We extend the current text-based language modelling approach for video shot retrieval (Westerveld, de Vries and van Ballegooij, 2003) by proposing hierarchical variations of the Dirichlet, Absolute interpolation and Witten-Bell smoothing methods for the hierarchical physical video structure (shot within adjacent shots within a video) and for the hierarchical semantic video structure (shot within adjacent shots within a story)

We take a different approach to visual retrieval than the continuous GMM language modelling approach (Westerveld, de Vries and van Ballegooij, 2003), where they model their single visual feature, multi-spectral DCT-based texture, using a Gaussian Mixture Model In our approach we model multiple visual features, such as colour, edge and texture, using discrete probability models that can be smoothed with the same standard language modelling smoothing methods as used for text We spend significant effort in evaluating this language modelling approach for the HSV colour, Canny edge and DCT texture features For each feature we evaluate multiple global and regional discrete representations for the language modelling and standard visual retrieval approaches to the video shot retrieval task

In this thesis, we combine results for the text and visual features, such as colour, edge and texture, using variations of data-fusion methods originally developed for combining the results of multiple text search engines, which consist of either combining normalised scores or normalised ranks of the individual search engine (Lee, 1997)

In our experiments we investigate many of the language modelling approaches to information retrieval for both the text and visual-based retrieval of video shots We contrast these language modelling approaches with a couple of hand picked non-language modelling approaches such as TF-IDF and BM25 for text retrieval, and Manhattan distance, Euclidean distance and Jensen-Shannon distance for visual retrieval However, for both text and visual results we are primarily interested in the language modelling results and therefore pursue and extend them more thoroughly than the standard visual or text-based information retrieval approaches

In this thesis we are concerned with the evaluation of a discrete language model approach to information retrieval for both text and visual-based search of relatively large video collections The advantage of our approach is that it provides a consistent, effective and efficient method of video retrieval It provides a consistent basis from which other researchers can extend using the wider language modelling framework for example to do relevance feedback or even other applications such as topic tracking and alerting in the video domain

In the following sections we will describe our proposed approach in more detail in section 2 we describe our extensions to the text-based hierarchical language model for

video retrieval, in section 3 we describe our visual-based discrete language modelling approach to video retrieval, in section 4 we describe the fusion methods to combine multiple features and visual examples, and in section 5 we outline our evaluation methodology We end this chapter with a description of related research

4.2 Proposed Extensions to the Text-based Hierarchical Language Model for Video Shot Retrieval

The structure of videos is hierarchical since we can view a shot as being in a sequence of adjacent shots (local context) and likewise these adjacent shots are within the larger video unit such as a programme We refer to this as the physical hierarchical structure because it uses relationships solely based on the physical editing composition of the video In television news programmes each news story is very distinct from the other stories within the same programme and therefore another hierarchical structure based on the semantic or logical structure of a news programme would be more appropriate For television news this semantic hierarchical structure does not involve the video unit and instead is based on a shot within a small sequence of shots within a story In our experiments in the next few chapters we evaluate the effectiveness of different indexing units such as shots, adjacent shots, enclosing story and enclosing video for representing shots and also different physical and semantic hierarchical structures in the language modelling approach to video shot retrieval

We can view the language modelling approach to information retrieval as having two distinct types of smoothing namely discounting-based and combination-based The combination-based smoothing methods Jelinek-Mercer, Absolute, Dirichlet and Witten-Bell smooth the document model by combining it with the background collection model The Jelinek-Mercer method, which uses a simple linear interpolation, was extended by (Westerveld, de Vries and van Ballegooij, 2003) for video shot retrieval so as to smooth the shot's text model with the adjacent shot's text and with the whole video's text as well as of course the collection text model In their research Westerveld, de Vries and van Ballegooij referred to this structure as *shot+scene+video* whereas we will refer to it as the *shot+adj+video* to differentiate it from our semantic structures Their hierarchical Jelinek-Mercer language model (or linear interpolation language model) is defined as

$$\begin{aligned}\mathbb{P}_{interp}(w|SHOT, \lambda_{shot}, \lambda_{adj}, \lambda_{vid}, \lambda_{col}) = \; & \lambda_{shot} \times \mathbb{P}(w|SHOT) \\ & + \lambda_{adj} \times \mathbb{P}(w|ADJ) \\ & + \lambda_{vid} \times \mathbb{P}(w|VID) + \lambda_{col} \times \mathbb{P}(w|COL)\end{aligned} \tag{89}$$

where $\mathbb{P}(w|SHOT)$, $\mathbb{P}(w|ADJ)$, $\mathbb{P}(w|VID)$ and $\mathbb{P}(w|COL)$ are respectively the maximum likelihood probability of a word in the shot, in the window of adjacent shot text, in the text from the enclosing video and in the collection text model The parameters λ_{shot}, λ_{adj}, λ_{vid} and λ_{col} are set by tuning on an appropriate collection This linear interpolation of the probability estimators can be viewed as an hierarchical combination where first the video probability model is smoothed with the collection model, then the adjacent shot model is smoothed with this smoothed video

model, and finally the shot model is smoothed with this smoothed adjacent shots model This interpretation of the combination as occurring at each stage in the hierarchy between the current level and the higher level leads us to propose extensions to this text-based hierarchical language modelling approach to video shot retrieval that combine the hierarchical video structure in a similar hierarchical ordering but with the other combination-based smoothing methods Witten-Bell, Absolute and Dirichlet We will now define each of these hierarchical smoothing methods for the *shot+adj+video* structure – simply replacing the video's text distribution with the story's text distribution leads to the similar definitions of these hierarchical smoothing methods for the semantic *shot+adj+story* structure The *shot+video* and *shot+story* hierarchical structures are also easily adapted from these by excluding the *adj* level

We define the hierarchical Witten-Bell smoothing for the hierarchical *shot+adj+video* structure as

$$\begin{aligned}\mathbb{P}_{hier\_wb}(w|\ SHOT) = \ & \mathbb{P}_{WB}(w|\ SHOT, \\ & \mathbb{P}_{WB}(w|\ ADJ, \\ & \mathbb{P}_{WB}(w|\ VID, \mathbb{P}_{ML}(w|\ COL))\ ))\end{aligned} \tag{90}$$

where $\mathbb{P}_{wb}(\ )$ is as defined for Witten-Bell smoothing in Equation 21 (Chapter 2, page 18) In this hierarchical Witten-Bell smoothing we first smooth the video with the collection model using the Witten-Bell smoothing function, we then smooth the adjacent shots with this smoothed video estimator using the Witten-Bell smoothing function, and finally we smooth the shot model with this smoothed adjacent shots model using the Witten-Bell smoothing function

Likewise, we define hierarchical Absolute smoothing as

$$\begin{aligned}\mathbb{P}_{hier\_abs}(w|\ SHOT, \delta_{shot}, \delta_{adj}, \delta_{vid}) = \ & \mathbb{P}_{ABS}(w|\ \delta_{shot},\ SHOT, \\ & \mathbb{P}_{ABS}(w|\ \delta_{adj},\ ADJ, \\ & \mathbb{P}_{ABS}(w|\ \delta_{vid},\ VID,\ \mathbb{P}_{ML}(w|\ COL))\ ))\end{aligned} \tag{91}$$

where $\mathbb{P}_{ABS}(\ )$ is as defined for Absolute interpolation smoothing in Equation 22 (Chapter 2, page 18)

And similarly, we define hierarchical Dirichlet smoothing as

$$\begin{aligned}\mathbb{P}_{hier\_dir}(w|\ SHOT, \mu_{adj}, \mu_{video}, \mu_{col}) = \ & \mathbb{P}_{DIR}(w|\ SHOT,\ \mu_{adj}, \\ & \mathbb{P}_{DIR}(w|\ ADJ,\ \mu_{vid}, \\ & \mathbb{P}_{DIR}(w|\ VID,\ \mu_{col},\ \mathbb{P}_{ML}(w|\ COL))\ ))\end{aligned} \tag{92}$$

where $\mathbb{P}_{DIR}(\ )$ is as defined for Dirichlet smoothing in Equation 20 (Chapter 2, page 18)

In the next chapter we evaluate these hierarchical smoothing methods on the TRECVid 2002, 2003 and 2004 collections We compare these methods with the original hierarchical text-based language-modelling approach to video retrieval (Westerveld, de Vries and van Ballegooij, 2003) and with standard information retrieval and language modelling approaches to text retrieval We compare these proposed hierarchical smoothing methods on the physical *shot+video* and *shot+adj+video* hierarchical structures and on the semantic *shot+story* and *shot+adj+story* hierarchical structures In the *shot+adj+story* hierarchical structure the window of adjacent shot text is bounded within the story unit and therefore does not add as much noise as in the *shot+adj+video* hierarchical structure We believe the hierarchical Dirichlet language model is

well-founded as it reduces the amount of smoothing for large amounts of text samples at each level However, in the physical hierarchical structure large amounts of text in the window of adjacent shot text may not indicate a more reliable estimate These smoothing methods may be more appropriate for the semantic structures due to the clearer topical connections between levels of the hierarchy

4.3 Proposed Visual Language Models

In this thesis, we propose a discrete language modelling approach for visual-based retrieval of video shots We extend the application of the text-based language modelling approaches to visual-based retrieval by converting the visual features, such as colour, edge and texture, into suitable discrete representations This is in contrast to the continuous language modelling approach (Westerveld, de Vries and van Ballegooij, 2003), which uses the Gaussian Mixture Model, a continuous probability model, to model the visual feature for the video retrieval task

We convert a multidimensional continuous feature into a discrete feature by quantising each continuous dimension into a fixed number of value ranges In some cases, such as for texture, we reduce the number of dimensions to only the most significant We can also apply this process to existing discrete multidimensional features that have ordinal dimensions in order to reduce the size of their indexing language After converting each dimension to a small set of discrete ranges for a fixed number of dimensions, we use the well established multidimensional histogram as our feature representation and we treat the values in each bin in the histogram as the count for a distinct symbol in the visual language for the particular feature

This interpretation of a visual feature as a discrete distribution of visual symbols is amenable to the same smoothing methods that have been used in the past for the text-based language modelling approaches to information retrieval It is our intention in this thesis to evaluate the originally text-based smoothing methods on our colour, edge and texture visual languages to validate our approach and to compare its effectiveness with standard visual retrieval models There is nothing particularly linguistic in the standard text-based language modelling approaches for information retrieval and therefore it is our belief that the text-based language modelling approach is also applicable to visual languages Since the visual features are less semantic than text features, different smoothing methods than in text retrieval may be more effective for some visual features or maybe even for all In particular, discounting methods may be sufficient for smoothing the maximum likelihood estimates If features have close to uniform distributions across the collection then there will be little difference between discounting and interpolation smoothing methods

We believe that discrete representations of non-semantic features that have symbols of near uniform distribution across the collection are more desirable as they can contain more information than the same features with a skewed discrete representation of the same language size The maximisation of the information carrying potential of a discrete language is especially important when the original continuous visual feature is extremely skewed This occurs for our DCT texture feature and if we uniformly quantised its range, we would create a visual language where most of the probability mass would be located in just a few discrete symbols for all documents

The benefits of our discrete visual language modelling approach is that it is consistent with the text-based language modelling approach and can directly avail of existing research, such as with regards to smoothing methods Since our feature representation is a multidimensional histogram we do not need to assume that the feature's statistical distribution follows a particular parametric probability density, such as a Gaussian Mixture Model Our discrete approach to feature representation is more efficient to index, since unlike the Gaussian Mixture Model there is no need to perform iterative EM parameter learning, and it is also more efficient for retrieval since the speed of query-likelihood for most retrieval models is related to the number of unique terms in the query, which in general is smaller in our approach due to the quantisation of the continuous features into a smaller number of distinct values Our approach is widely applicable to many of the effective visual feature representations that are used in current video retrieval systems such as histograms, correlograms and co-occurrence matrices of colour, edge and texture features This visual language modelling approach can also be applied to standard MPEG7 descriptors such as MPEG7 Scalable Colour, MPEG7 Colour Structure, MPEG7 Edge Histogram and MPEG7 GoF/GoP Colour descriptors We will also show later in this thesis that the visual language modelling approach is generally as effective as traditional visual retrieval models In summary, the visual discrete language modelling approach to video retrieval is consistent with text retrieval, efficient, effective and widely applicable to many visual features It puts text and visual retrieval in the same discrete language modelling framework allowing for easier extension of this research across media such as for relevance feedback The major weakness in our approach is that it cannot support high-dimensional feature representations as the storage requirements grow exponentially with the number of dimensions

On the other hand the benefits of the Gaussian Mixture Model (GMM) language modelling approach (Westerveld, de Vries and van Ballegooij, 2003) are that the indexed representation of the feature is smaller than for the discrete histogram approach and it can handle higher-dimensional features since its storage requirements grow only linearly (diagonal covariance) or squared (full covariance) with the number of dimensions The samples also guide the representation through the EM parameter learning procedure and this can be interpreted as wasting less information in the representation on insignificant or non-existent patterns than might be the case for the histogram approach The weaknesses of this approach is that it is slower to index due to the required EM training and is also slower to retrieve from due to each sample being distinct and since the GMM is slower to evaluate than looking up the probability in a multidimensional histogram The assumption that features can be represented as a GMM may also limit its applicability to other types of features that are not as well approximated by this distribution The use of a fixed number of components in the GMM may also limit its effectiveness for features that have very varied amounts of complexity in each shot

We apply our discrete visual language modelling approach to the colour, edge, and texture features, which we chose to be as independent of each other as possible, in the hope that when we combine their results they will complement each other, though there is possibly some overlap in the information content between texture and colour and also between texture and edge For each of these feature classes we have chosen a single feature instance for colour we chose HSV colour, which has been successfully used in many image search engines (Faloutsos et al , 1994, Hauptmann et al , 2004), for edge we chose the Canny edge detector, which has been used in (Hauptmann et al , 2004) for visual retrieval and for texture we chose DCT coefficients which have been used by (Vasconcelos and Lippman, 2000, Westerveld, de Vries and van Ballegooij,

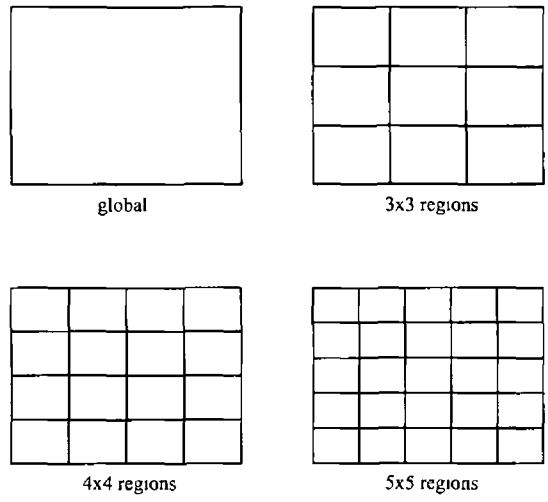


Figure 20 Global and regional whole image representations

2003) for image and video retrieval Our approach is directly applicable to other visual features that can be represented as a distribution of single-dimensional or multidimensional numeric or non-numeric samples

For each feature we compare multiple global and regional discrete representations for the language modelling and standard visual retrieval approaches for the video shot retrieval task We construct regional versions of our features by including their relative X and Y position in the keyframe, scaled between 0 and 1, as part of the feature's multidimensional representation We compare the retrieval performance for each feature using this regional representation quantised into 3x3, 4x4, and 5x5 regions per image as illustrated in Figure 20 These representations provide a limited but still potentially beneficial amount of regional position information However, in contrast to a full visual region retrieval system, we perform only whole image querying and do not support querying or matching based on a part or region of an image The global feature representations test the visual language modelling approach with small dense visual languages, whereas the regional feature representations test the visual language modelling approach with large sparse visual languages

We will now describe our colour, edge and texture visual languages

4 3 1 Visual Colour Language

We choose the HSV colourspace for the visual colour language due to its relative success in other video retrieval systems (Faloutsos et al , 1994, Hauptmann et al , 2004) The approach we take for colour retrieval is equally applicable to other colourspaces such as the MPEG-1 encoded YCbCr colourspace or the LAB colourspace that is based on human perception of colour differences However, in this thesis we will investigate a single feature instance for each visual feature class and will leave it to future work to broaden the evaluation of our approach to the many other types of colour, edge or texture

For the HSV colour feature we assume that each pixel's colour information is independent of the surrounding pixels This assumption, which is not strictly true, permits us to model an

image in terms of independent and identically distributed (i i d) colour pixel samples The positional colour feature has five dimensions X, Y, H, S, V in which the X and Y dimensions are scaled from zero to one For the *global* colour representations we ignore the first two dimensions and quantise the H, S, V dimension, whereas for the *regional* colour representations we quantise all five dimensions As with all our visual features, we use a multidimensional histogram to store frequency counts of the quantised feature, but interpret this structure in terms of counts of distinct symbols in our visual language model for the feature

In our global HSV experiments, we quantise the HSV colour feature into different histogram representations *H 80+1*, *HSV 5x5x5* and *HSV 16x4x4* The first representation *H 80+1* quantises a HSV value into 80 uniform levels (histogram bins of equal width) based on hue alone with unsaturated colours, which have unreliable hue values, stored in a separate out-of-bounds bin This represents a very simple single dimensional representation of HSV colour, which results in a language of 81 symbols In contrast, the other representations that we evaluate quantise the three colour bands - hue, saturation, and value (brightness) The first of these multidimensional representations is *HSV 5x5x5* which quantises each dimension uniformly into 5 levels This representation is the same as CMU's TRECVid 2003 colour histogram representation (Hauptmann et al , 2004), which they also use in a 5x5 regional representation, and it produces a language of 125 symbols for its global representation The final representation we evaluate is *HSV 16x4x4*, which is similar to the MPEG7 Scalable Colour descriptor (MPEG7 Committee, 2002), in which each band is uniformly quantised but with the hue band given more levels, 16 compared to 4 each for the other two bands, and this representation produces a language of 256 symbols We evaluate these different global HSV colour representations in order to determine the overall best HSV representation for the TRECVid search task and to compare visual language models and standard retrieval models

After establishing the best global colour representation we investigate its use for regional colour representations In our regional HSV colour representation we quantise the X, Y position dimensions for each HSV sample uniformly into 3, 4 or 5 levels, thereby essentially breaking the image into 9, 16, or 25 independent rectangular regions and of course producing a visual language with 9, 16, or 25 times the number of symbols as in the global representation In all our experiments on these visual languages we investigate discounting-based and interpolation-based smoothing methods for the discrete visual language modelling approach as well as other traditional visual histogram retrieval models such as Manhattan distance, Euclidean distance and Jensen-Shannon distance

4 3 2 Visual Edge Language

As described in the previous chapter, an edge is an abrupt change in image intensity and usually has both an associated direction and magnitude and is located at a specific pixel within the image There are many edge detectors, such as based on the Laplacian, Prewitt, Sobel, Robinson, or Kirsch operators (Sonka et al , 1998), however we will restrict our evaluation to the Canny edge detector which was previously used for video shot retrieval in (Hauptmann et al , 2004)

In our Canny edge feature, we ignore the magnitude of the edge and constrain the edge direction to an 180° range with the first quantisation level of our discrete representation of edge

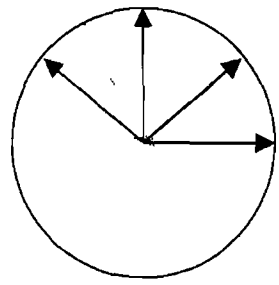


Figure 21 Canny edge feature for 4 directions with decision boundaries indicated with dashed lines

direction centred on the horizonal axis as illustrated in Figure 21 We treat non-edge samples as having an out-of-bounds edge direction which requires an additional symbol for each visual edge representation Similar to colour, we assume that the edge features, which exist for each pixel in an image, are i i d and therefore can be modelled as independent samples in a distribution

We first compare the global visual edge languages *Canny 4+1*, *Canny 16+1*, *Canny 32+1* and *Canny 64+1* which have from 4 to 64 levels of quantisation for the edge direction and with an extra symbol for the out-of-bounds edge direction, producing a language consisting of between 5 and 65 symbols After investigating retrieval models for these global visual edge languages and establishing a reasonable global representation of edges, we investigate extensions to this visual language for 3x3, 4x4 and 5x5 regions The non-edge symbols, which typically contain most of the probability mass of this distribution, makes this feature very different from the more uniform distribution of the HSV colour feature We would expect that interpolation-based language models would be more beneficial than discounting-based language models for this feature as they reduce the importance of common symbols in the calculation of query-likelihood

4 3 3 Visual Texture Language

In this section we present our visual texture languages Texture can be simply defined for images as patterns in a local spatial neighbourhood and similar to (Westerveld, de Vries and van Ballegooij, 2003) we use the DCT coefficients as our texture feature Whereas Westerveld, de Vries and van Ballegooij (2003) use multi-spectral DCT coefficients, we limit our representation to DCT coefficients based on the luminance band, the Y band of YCbCr colourspace, disregarding the DCT coefficients based on the colour bands Cb and Cr The luminance DCT coefficients are calculated from non-overlapping 8x8 pixel blocks in MPEG-1 and JPEG encoded images and they can be efficiently extracted from encoded streams without a full decode To simplify our software implementation (programming effort), we fully decode the keyframes using standard image libraries and we then perform the DCT transform to produce the DCT coefficients for our experiments instead of writing our indexing software to do a more efficient partial decode of the JPEG and MPEG-1 streams

The first DCT coefficient is the mean luminance of the 8x8 pixel block while the other coefficients each represent the amplitude of Cosine waves of specific frequencies (see Section 3 2 4 2, page 59) The Cosine amplitudes are unbounded values that can be positive or negative and are ordered from low to high frequencies where the first few coefficients, the low frequencies, contain most of the important visual information for encoding an image In fact JPEG and

MPEG-1 encoding standards use this principle when compressing an image by concentrating the majority of encoded bits to representing only the low frequency DCT coefficients We apply the same focus on the low frequency coefficients in our representation of texture for our experiments

In minor ways, our DCT texture representation overlaps the previous two features, colour and edge, since the first DCT coefficient is the average brightness and since edges are also a spatial feature However, the colour and edge features are more focused on their specific feature type and are also at a finer pixel level than the DCT texture representation, which is calculated for non-overlapping 8x8 pixel blocks as opposed to individual pixels We expect that these three features should complement each other when combined into a single retrieval model In contrast, in the approach of (Westerveld, de Vries and van Ballegooij, 2003, Vasconcelos and Lippman, 2000) multi-spectral DCT coefficients constitute their full visual-based retrieval model

In our experiments we represent texture by using different numbers of significant DCT coefficients into a fixed number of quantisation levels Similar to previously described features, our DCT texture representation is essentially a multidimensional histogram In contrast, in the approach of (Westerveld, de Vries and van Ballegooij, 2003, Vasconcelos and Lippman, 2000) the DCT probability model is a multivariate Gaussian Mixture Model, which is a more compact representation that can more easily model larger numbers of DCT coefficients or bands into a single representation

Each of our DCT texture representations quantise a fixed number of low frequency DCT coefficients into a fixed number of bins per coefficient The quantisation levels, boundaries between bins, for each dimension of the histogram are calculated across the whole keyframe collection so that the marginal distribution of a specific DCT coefficient uniformly populates its quantisation bins For example, for 10 bins representing a DCT coefficient, one tenth of the values of the DCT coefficient will populate each bin in its marginal distribution for the collection of all shot keyframes In other words, we use variable width bins in contrast to our edge and colour representations where the bins in each dimension had the same width Since the Cosine amplitudes are essentially unbounded from negative infinity to positive infinity with most amplitudes centred around the zero value, a uniform partitioning of the Cosine wave amplitude's range would attribute the same symbol to many texture representations, which would as a result reduce the discrimination power of the indexing language

In our experiments on the TRECVid collections we compare the global texture languages that use between 2 and 5 DCT coefficients, specifically the *DCT 10x10*, *DCT 8x8x8*, *DCT 4x4x4x4*, and *DCT 3x3x3x3x3* visual texture languages, which consist of 100, 256, 256 and 243 symbols respectively We will first investigate texture as a global feature before experimenting with it for regional texture representations of 3x3, 4x4 and 5x5 regions

4.4 Fusion Methods

In this thesis we investigate the fusion of retrieval models in order to combine (A) the multiple visual features, (B) the multiple visual examples and (C) the multiple modalities text and visual as illustrated in Figure 22 The combination (A) supports the retrieval of video shots using a single visual example and involves the automatic fusion of colour, edge and texture

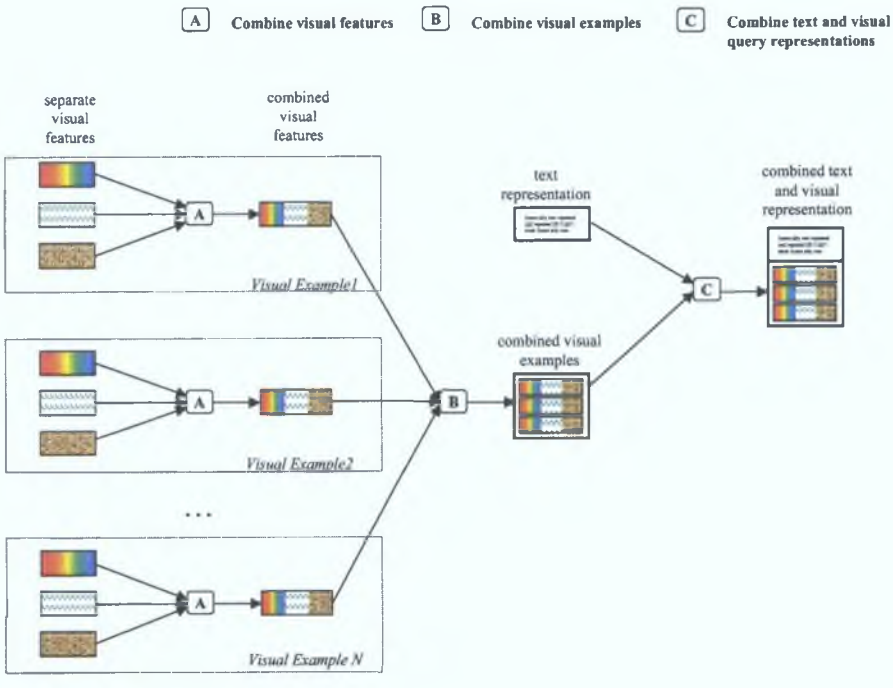


Figure 22: Fusion of individual retrieval models for TRECVid topics.

retrieval models. The combination (B) supports visual-based retrieval of video shots using a query with multiple visual examples and involves the automatic fusion of results from possibly quite disparate image or video examples. While the combination (C) supports the retrieval of video shots using a query with both text and multiple visual examples for which the combination would involve very different result sets.

As discussed in the previous chapter, there is a lack of empirical studies of fusion techniques in video retrieval and it is not clear what techniques are best for the different fusion tasks required by video retrieval systems and in particular what fusion methods are best for our visual language modelling approach. We investigate late fusion methods originally developed for combining the results of multiple text search engines (Lee, 1997; Fox and Shaw, 1994). We compare fusion methods based on normalised score and normalised rank which use either the average, weighted average or maximum of individual results as the combination function. The weighted average combination is particularly important in video retrieval as we believe that some features are in general better than others though this could also be used in text retrieval to prefer the results of particular search engines over others. We also compare these results with a single probabilistic combination that assumes all features and examples are fully independent.

We use the following notation to refer to each fusion strategy:

- *CombJointPr* - multiply the probabilities of individual generative models (or add the log-likelihoods).
- *CombSumScore* - add the normalised scores of the results (traditional CombSUM fusion)
- *CombSumRank* - add the normalised ranks of the results (Borda count).

- *CombMaxScore* - order by the maximum of the normalised scores of the top results
- *CombMaxRank* - order by decreasing normalised rank removing duplicates (equivalent to round-robin when result lists are the same size)
- *CombSumWtScore* - weighted average of the normalised scores of the results (linear interpolation)
- *CombSumWtRank* - weighted average of the normalised ranks of the results (weighted Borda count)

As in (Lee, 1997) we define normalised rank as

$$nrank_{doc} = \frac{N - rank_{doc}}{N}, \tag{93}$$

and normalised score as

$$nscore_{doc} = \frac{score_{doc} - score_{min}}{score_{max} - score_{min}}, \tag{94}$$

where $score_{min}$ is the score at document rank $N+1$ and all documents not in the top N results are given a score of zero for both the normalised rank and normalised score The parameter N controls the truncation of the original result list thereby removing the noise that may be generated from low-ranking document scores

We use log-likelihoods as our score for each document in our text and visual language model's retrieval results since the visual features' generative probabilities are very small and cannot be directly represented using double precision floating point numbers As a result we are limited in how we can efficiently combine the probabilities but one simple combined generative model, which we also evaluate, is to assume that all the features and visual examples are independent, which is straightforward to calculate by adding the log-probabilities

4 4 1 Combining Multiple Features

For the first fusion task (A) to combine the multiple visual features colour, edge and texture we compare the following fusion methods *Vis-CombJointPr*, *Vis-CombSumWtScore* and *Vis-CombSumWtRank* methods

The *Vis-CombJointPr* is a generative ranking model for the three features and is defined as

$$\mathbb{P}(col \wedge edge \wedge texture|shot) = \mathbb{P}(col|shot) \times \mathbb{P}(edge|shot) \times \mathbb{P}(texture|shot), \tag{95}$$

where each visual feature is assumed to be independent This method assumes each feature to be equally important and will probably allow the noise from the poor-matching feature's results or from the low-ranking results to overwhelm this joint probability A better strategy could be to normalise all low ranking probabilities to a small constant background probability The probability approach may be more competitive with other approaches if it were efficient to calculate a finite mixture model (linear interpolation) of the features generative probabilities, such as

$$\begin{aligned}\mathbb{P}(colour \wedge edge \wedge texture|shot) =& \lambda_{colour} \times \mathbb{P}(colour|shot) + \lambda_{edge} \times \mathbb{P}(edge|shot) \\ &+ \lambda_{texture} \times \mathbb{P}(texture|shot)\end{aligned} \tag{96}$$

Since the event spaces for the components of this finite mixture model are not equivalent, this approach is not as elegant as when the finite mixture models are used for smoothing in the language modelling approach to information retrieval

The closest we get to using the generative probabilities in our other fusion strategies is with the *Vis-CombSumWtScore* method which combines the normalised scores (log-likelihoods) of the results of each feature, and is defined as

$$score = w_{colour} * nscore_{colour} + w_{edge} * nscore_{edge} + w_{texture} * nscore_{texture}, \quad (97)$$

where $w_{colour} + w_{edge} + w_{texture} = 1$

The fusion method *Vis-CombSumWtRank* ignores the scores (log-likelihoods) completely and combines simply based on the average weighted normalised rank across features, and is defined as

$$score = w_{colour} * nrank_{col} + w_{edge} * nrank_{edge} + w_{texture} * nrank_{texture}, \quad (98)$$

where $w_{colour} + w_{edge} + w_{texture} = 1$

We make no assumptions about the weights in these parametric fusion methods *Vis-CombSumWtScore* and *Vis-CombSumWtRank* but instead train and test them on independent collections In our results, we present both optimised and unbiased results for these fusion models We set the size (N) of the truncated result lists to 1000 for both fusion methods instead of tuning on a separate collection

We would prefer the normalised score method over the normalised rank fusion method for the simple reason that it makes more use, if only indirectly, of the query-likelihood Varying the amount of results combined from each feature may have been beneficial, however since we are primarily interested in comparing fusion methods, we did not do this as it would only further complicate the interpretation of the results We intend to look into variable result sizes in future work Another strategy that may be beneficial would be to try curve-based rank normalisation functions rather than the straight line that we currently use for normalising the rank score An more interesting extension to this work would be to combine the results using score distributions (Manmatha et al , 2001)

4 4 2 Combining Multiple Visual Examples

For the second fusion task (B) to combine multiple visual examples, we compare the following fusion methods *VisExs-CombJointPr*, *VisExs-CombSumScore*, *VisExs-CombSumRank*, *VisExs-CombMaxScore* and *VisExs-CombMaxRank*

For the *VisExs-CombJointPr* we will use the previously described *Vis-CombJointPr* for the probability of each separate visual example This fusion method creates a generative model under the assumption of independence between visual examples This is actually equivalent to assuming that all samples from all visual examples are sampled independently from the same single source due to our earlier assumption of independence of the features' samples

For the other fusion methods we can combine the results of any of the previously described visual feature combinations For the *VisExs-CombScore*, *VisExs-CombRank* and *VisExs-CombMaxScore* methods we will normalise a multiple (i e $scale * (1000/number\ of\ examples)$) of the amount of documents required from each visual example in order to get a result size of at least 1000 We select the value of *scale* by the usual method in this thesis of optimising on one test collection and validating on another In general the best values of *scale* are between 1 05 and 1 20 though sometimes it reaches 2 0 A better approach may be to set this to some fixed value for all the test collections

The *VisExs-CombJointPr* fusion method will probably suffer from low-ranking shots in each result affecting the good results of the visual examples The *VisExs-CombSumScore* (traditional CombSUM) and *VisExs-CombSumRank* (Borda count) combine the results for each example in a way that allows the top ranking results to affect each other in the assumption that this will improve the ranking by reinforcing the positive results and reducing the noise These fusion methods require that relevant results from the source result lists overlap more so than non-relevant documents in order to have successful fusion The *VisExs-CombMaxRank* take a strictly independent view of multiple examples and orders the results in a round-robin fashion The assumption is that the separate results for each visual example do not strongly overlap, and if combined by averaging score or rank would reinforce the noise more than they reinforce the common relevant results The *VisExs-CombMaxScore* again assumes that the results do not overlap in a way that would be useful but assumes that the normalised scores hold some useful relevance information that can be used to measure the separate results in a better manner than round-robin

We will evaluate these fusion methods using a fully automatic evaluation procedure on the TRECVid 2002, 2003 and 2004 search tasks, the results of which represent a purely visual-based approach to the TRECVid search task

4 4 3 Combining Text and Visual

For the third fusion task (C) to combine the results from text and visual models we compare *TextVis-CombJointPr*, *TextVis-CombSumWtScore* and *TextVis-CombSumWtRank* fusion methods

TextVis-CombJointPr is a generative fusion method which combines the probability of the shot producing the text of the query with the probability of the shot generating the different visual examples under the assumption that both the text and visual models are independent The same potential problems apply to this combination as to all *CombJointPr* combinations except that for this combination the reduction in performance due to added noise is likely to be more severe since the relatively good text results are combined on equal terms with the relatively poor visual results, which will most likely achieve significantly lower results than for text alone The visual generative probabilities are for thousands of pixel samples and will likely overwhelm the text generative probabilities which are only for a couple of query text terms

The *TextVis-CombSumWtScore* method combines the normalised scores of the top 1000

results of the text and visual results and is defined as

$$score = w_{text} * nscore_{text} + w_{visual} * nscore_{visual}, \quad (99)$$

where $w_{text} + w_{visual} = 1$ The fusion method *TextVis-CombSumWtRank* (weighted Borda count) simply combines the text and visual results based on the weighted average of the normalised rank across features, and is defined as

$$score = w_{text} * nrank_{text} + w_{visual} * nrank_{visual}, \quad (100)$$

where $w_{text} + w_{visual} = 1$

As before, we make no assumptions about the weights in the parametric fusion methods *TextVis-CombSumWtScore* and *TextVis-CombSumWtRank* but instead train and test them on independent collections

We will evaluate these fusion methods using a fully automatic evaluation procedure on the TRECVid search tasks and these results represent a combined text and visual-based approach to the TRECVid search task We expect that either the *TextVis-CombSumWtScore* or *TextVis-CombSumWtRank* will perform best and that the *TextVis-CombJointPr* will perform very poorly We are not assured of getting better results for any of the combined text and visual approaches than for the text approach alone because positive improvements in some topics may be overshadowed by the negative effects the visual results have on other topics

We will also evaluate the combination of text and visual features for two ideal situations In the first situation *TextVisBoth-Oracle* we assume that an Oracle selects whether the topic will be text only, visual only or combined text and visual In the second variation *TextVisComb-Oracle* we allow the Oracle to select the weights for combining text and visual results on a per-topic basis In both variations the Oracle selects the best of the available choices The purpose of the *TextVisBoth-Oracle* is to achieve an estimate of the performance of our approach if an ideal user was able to make a simple but always correct choice for each topic between using the text, the visual or both for each query The purpose of the *TextVisComb-Oracle* is to identify the best the system (or a user) could achieve if it modified the weights of text and visual features on a per topic basis These experiments will give us some indication of the performance potential for combining visual retrieval with text retrieval for the video search task A related fusion approach is to treat each topic as being a member of a query-class and use class-dependent weights for the fusion task (Yan et al , 2004) We do not pursue this approach but believe it would be beneficial with the appropriate amount of training data

4.5 Evaluation Methodology

In this thesis, we show through a series of experiments for text-based, visual-based and combined text and visual searches that the discrete language modelling approach to text and visual based searching of video content produces competitive results compared to other approaches We evaluate our proposed extensions to the text-based hierarchical language modelling approach, our discrete language modelling approach to visual retrieval, and the effectiveness of the combined text and visual retrieval approach using fully automatic experiments on the TRECVid 2002, 2003 and 2004 search tasks

The TRECVid 2002 and TRECVid 2003 collections are very different and as a result we sometimes get inconsistent results from both collections for certain features or representations The TRECVid 2002 test collection contains old video content, public information and commercial programmes predominantly from the 1950's onwards of a mixed but generally poor visual quality, while the TRECVid 2003 and TRECVid 2004 collections contain ABC and CNN news from the late 1990's of a higher and consistent visual quality Each TRECVid collection is supplied with a common automatic speech recognition (ASR) text, shot definitions, keyframes, search topics and relevance judgements which we use in our experiments The TRECVid search topics, which are listed in the appendices, were composed by NIST to represent 'real-world' visually-oriented information needs concerning people, things, events and locations, and each TRECVid topic consists of a text description and multiple video and image examples

We perform fully automatic experiments where the retrieval system takes the TRECVid topic without modification by a user Our system may then modify it, such as performing stopword removal and stemming for the text description, but all modifications of the topic are performed automatically without user intervention This is in contrast to the manual search procedure more popularly employed in TRECVid, which allows a user to translate the TRECVid topics into a form more suitable for each search system We believe that fully automatic experiments, which potentially decrease the reported performance compared to manual TRECVid experiments, provide a more unbiased view of system performance that can be compared more easily and repeated independently within the research community In the TRECVid 2004 conference the submission of fully automatic runs were provisionally supported to test their usefulness for their potential adoption in the TRECVid 2005 evaluation

For all our experiments we will report both the unbiased results and the optimised results for each retrieval model that requires parameters to be set We choose the unbiased parameters through training the retrieval model on a separate collection in order to maximise the mean average precision (MAP) measure The biased results such as the optimised MAP in an experiment will be presented in brackets to distinguish them from the other experimentally valid results that one would expect to achieve in a blind unbiased test In our presentation of results we will primarily discuss the MAP measure but we will present mean precision at 10, 30, and 100 document cut-offs when space permits, as this correlates with what a user would perceive when looking at a short detailed list of the top 10 results, a more careful detailed look at the top 30 results and for example a compact image-only results layout of the top 100 keyframes

We evaluate our proposed approach to video shot retrieval in the next three chapters, which deal in turn with text-based retrieval, single visual feature-based retrieval and combined text and visual retrieval of video shots for the TRECVid search tasks

In Chapter 5 we investigate video shot retrieval using text evidence alone We compare standard language modelling and standard information retrieval approaches such as TF-IDF and BM25 for different indexing units such as shots, adjacent shots, videos and stories for the TRECVid search tasks in order to create a baseline for non-structural approaches to video retrieval We also compare the hierarchical language modelling approach (Westerveld, de Vries and van Ballegooij, 2003) with our proposed extensions that use either hierarchical Absolute, hierarchical Dirichlet or hierarchical Witten-Bell smoothing on the hierarchical physical *shot+video* and *shot+adj+video* structures and the hierarchical semantic *shot+story* and *shot+adj+story*

structures

In Chapter 6 we evaluate our discrete language modelling approach for video shot retrieval on the separate colour, edge and texture visual features We compare the language modelling approaches to other standard visual retrieval ranking models, such as Manhattan distance, Euclidean distance and Jensen-Shannon distance, for individual visual features and we establish the relative performance of different visual features on the TRECVid collections For each feature, we first present experiments on *global* representations of the feature in which we do not take any positional information into account After establishing results for the global representation we experiment with representations that take into account positional information for 3x3, 4x4 and 5x5 regions

In Chapter 7 we evaluate the fusion of results from multiple features, multiple visual examples and multiple modalities The key problem addressed in this chapter is not which text retrieval model achieves the best fusion results but how to combine the visual features and visual examples successfully with each other and with the text results We experiment with combining features from a wide selection of visual retrieval models, discounting-based language models, interpolated-based language models, and standard visual retrieval models, in order to achieve a wider and hopefully more balanced view of the benefits and faults of different fusion methods

4.6 Related Research

We have published some partial comparisons of the discrete visual language modelling approach that is presented in this thesis In the TRECVid 2004 workshop we submitted retrieval runs for the automatic video search task using Lidstone smoothed and Jelinek-Mercer smoothed visual discrete language models with the regional HSV colour, Canny edge and DCT features (Cooke et al , 2005) Our choice of fusion strategy for our visual features in these submitted results was poor and only achieved a MAP of 0 018 for visual-only searching, though this was slightly better than the continuous GMM DCT visual language modelling approach, which achieved a MAP of 0 008 for their static model and 0 0010 for their dynamic model (Ianeva et al , 2005) Our submitted TRECVid 2004 fused text and visual results were actually the best of the submitted runs for the automatic video search task The Informedia (Hauptmann et al , 2005) and Lowlands (Ianeva et al , 2005) (hierarchical text language model with continuous DCT GMM visual language model) submitted automatic runs were only marginally lower – certainly not significantly different and they could easily have been better in terms of MAP considering the somewhat unpredictable nature of text and visual fusion We view these comparisons and the results in subsequent chapters as indicating that our approach is at least as good as the current state-of-the-art for this automatic retrieval experiment More recently, we compared fusion methods for the discrete visual Jelinek-Mercer smoothed language model on the different video retrieval fusion tasks (Mc Donald and Smeaton, 2005)

The research by the Lowlands group at TRECVid (Baan et al , 2002, Westerveld, de Vries and van Ballegooij, 2003, Westerveld et al , 2004, Ianeva et al , 2005) was a major inspiration for our research Their approach to visual video shot retrieval is to apply the Jelinek-Mercer smoothing method to a continuous GMM query-likelihood for the DCT feature Our research differs as it explores the discrete visual language modelling approach for visual features and

takes a multi-feature approach to visual video retrieval Our approach is quicker to index and retrieve due to its discrete nature and as we show in this thesis is as effective if not better For text-based video retrieval, we borrow their hierarchical language modelling approach but explore alternative hierarchical smoothing methods and alternative hierarchical video structures

The probabilistic source model (Jin and Hauptmann, 2002) is perhaps the first discrete visual language modelling approach for video retrieval, though it was not proposed in such terms It differs from our approach in that it uses document-likelihood and therefore possesses the previously discussed problems concerning ideal document match In their approach they apply their probabilistic source model to the Munsell colour of the keyframes with grid-based positional information but make questionable independence assumptions between dimensions of their feature representation They use a specific Dirichlet smoothing configuration to smooth their probability estimates that is equivalent to Laplace smoothing Our discrete visual language modelling approach more clearly has its roots in the language modelling approach and uses query-likelihood to rank documents It does not make over-simplifying independence assumptions between dimensions of our colour representation We also apply our discrete visual language modelling approach to other features such as Canny edge and DCT texture and investigate a wide range of discounting and interpolation smoothing methods derived from language modelling research

Parallel to our TRECVid 2004 work, de Vries and Westerveld (2004) compared their GMM approach with a discrete Jelinek-Mercer language modelling approach for the multi-spectral DCT feature Their findings were that the continuous DCT GMM approach is superior to the discrete Jelinek-Mercer language modelling approach They reported results indicating that the GMM approach is many times better than the discrete approach for the same multi-spectral DCT feature – MAP 0 0044 for discrete language model compared to MAP 0 0281 for continuous GMM on the TRECVid 2003 collection In their comparison the discrete visual language did not contain positional information but their continuous GMM representation did, which likely accounts for a significant amount of their reported performance difference Contributing factors to the poor performance of their discrete language model is that they quantised the DCT coefficients using a uniform partitioning and the high-dimensionality of their feature may have led to some estimation problems as well as reducing the potential speed benefit of a smaller discrete representation The parameters of the discrete language model were not chosen by tuning on a separate collection, which also limits the fairness of their test Our findings and discrete visual language for this feature differ Firstly, by using the marginal distribution of the DCT dimensions across the collection when selecting boundaries for the quantisation levels, we increase the discrimination power of our discrete language and secondly, by using smaller discrete languages we produce a quicker retrieval model with a more dense feature Our research is wider in scope as it investigates multiple visual features, multiple feature representations and different fusion strategies We also attempt a more careful quantisation of the DCT coefficients and compare our results to alternative similarity models for discrete feature representations such as L1, L2 and Jensen-Shannon distance

4.7 Summary

In this chapter we proposed our discrete visual language modelling approach for video retrieval. We extended the application of the text-based language modelling approaches to visual-based retrieval by converting the visual features, such as HSV colour, Canny edge and DCT texture, into suitable discrete representations. Our discrete visual language modelling approach can be further applied to many of the effective visual feature representations that are used in current video retrieval systems such as histograms, correlograms and co-occurrence matrices of colour, edge and texture features. This visual language modelling approach can also be applied to some of the standard MPEG7 descriptors.

We extended the current text-based hierarchical language modelling approach to video shot retrieval (Westerveld, de Vries and van Ballegooij, 2003) by considering hierarchical versions of the Dirichlet, Absolute and Witten-Bell language models for different physical and semantic video structures.

We proposed to combine results for the text and visual discrete language models using variations of data-fusion methods originally developed for combining the results of multiple text search engines. These fusion methods involve either combining results using normalised scores or normalised ranks with the average, weighted average or maximum combination functions. We also proposed to evaluate a generative fusion method that uses the joint probability of the features, visual examples and modalities under a feature independence assumption.

Additionally, we outlined our evaluation methodology for our proposed retrieval models, visual features and fusion methods, which involves the evaluation of our approach using fully automatic retrieval experiments on the TRECVid 2002, 2003 and 2004 search tasks. In the next three chapters we will evaluate this approach, first for text-only retrieval, then for single visual feature retrieval and finally for combined text and visual retrieval.

CHAPTER V

EVALUATION I VIDEO RETRIEVAL USING TEXT FEATURES

In this chapter we evaluate language modelling approaches for the text-based video shot retrieval task on the TRECVid 2002, 2003 and 2004 collections We compare standard retrieval models, discounting-based language models and combination-based language models using a shot representation of either the shot text, the video text, the story text or the adjacent text We also compare our proposed hierarchical Absolute, hierarchical Dirichlet and hierarchical Witten-Bell language models with the hierarchical Jelinek-Mercer language model using physical and semantic structures of video content Our purpose in this chapter is to establish a baseline for language models and standard retrieval models for the video shot retrieval task, to establish the best physical and semantic hierarchical video structures and to evaluate our proposed hierarchical language models

5 1 Introduction

5 2 Experiments with non-hierarchical structures

5 3 Experiments with hierarchical physical video structures

5 4 Experiments with hierarchical semantic video structures

5 5 Summary

5 1 Introduction

In this chapter we evaluate language modelling approaches for text-based video shot retrieval on the TRECVid collections We compare language models and standard information retrieval models for the non-hierarchical physical structures *shot-only*, *adj-only* and *video-only* and for the semantic *story-only* structure We evaluate hierarchical document structures for the video shot retrieval task which include both the physical structures *shot+video* and *shot+adj+video* and the semantic structures *shot+story* and *shot+adj+story* We also compare our hierarchical language models hierarchical Absolute, hierarchical Dirichlet and hierarchical Witten-Bell, which we proposed in the previous chapter, with the hierarchical Jelinek-Mercer (linear interpolation) language model (Westerveld, de Vries and van Ballegooij, 2003) for the physical and semantic hierarchical representations The purpose of this chapter is to identify the best physical and semantic shot representations and the best smoothing function for the text-based language modelling approach to video shot retrieval

We use a fully-automatic experimental setup, as described in the previous chapter, for our empirical evaluation of the retrieval models on the TRECVid 2002, TRECVid 2003 and TRECVid 2004 collections The official TRECVid manual search experiments have a similar setup except TRECVid manual search permits a professional user to translate the TRECVid

Table 6 Additional query stopwords for the TRECVid search topics

Additional query stopwords
find, additional, shots, scenes, pictures, containing, including, showing, lots, groups, multiple, partly, partially, visible

topics into a query representation suitable for the video retrieval system This in effect allows the reformulation of the text associated with a topic and hinders the comparison of results across research sites that use different query text The results presented here can differ from those of other published papers for the same algorithm when the original topic text has been manually modified or possibly stopped and stemmed differently

The objective of our experiments is not to find which parameter values for the different retrieval models provide the best possible performance but to identify which retrieval models provide transferable results across test collections We try to achieve these unbiased results, by optimising the retrieval models on one collection, and testing on a separate collection For example, the three BM25 parameters (k_1, k_2, b) are optimised on one collection in terms of MAP and then these parameter values are used to test the BM25 retrieval model on another collection The reported unbiased results for TRECVid 2002 use TRECVid 2003's optimised parameters, the unbiased results for TRECVid 2003 use TRECVid 2002's optimised parameters, while the unbiased results for TRECVid 2004 use TRECVid 2003's optimised parameters In our discussion of our results we are primarily interested in the unbiased MAP though we will also present performance measures for the optimised retrieval models

As discussed in the previous chapters the TRECVid 2002 collection is different from the TRECVid 2003 and TRECVid 2004 television news collections in terms of content, format and visual quality but the topics for all three collections are fairly similar representing information needs for locating video of persons, places, events and objects The ASR or audio somewhat guided the selection of topics for the TRECVid 2002 collection but more care was taken for the TRECVid 2003 and TRECVid 2004 collections not to allow the ASR to guide the formulation of the topic text The three TRECVid collection are provided with a common definition of all shots, the retrieval documents, which includes their temporal boundaries, their representative keyframes and importantly for this chapter a common ASR text transcript provided by LIMSI (Gauvain et al , 2002a) The aligned story boundaries for the ABC and CNN news programmes in the TRECVid 2003 search test collection were provided post TRECVid and we use these to define our semantic story representation The aligned news stories do not exist for the six C-SPAN programmes and for two of the ABC news programmes in the TRECVid 2003 search test collection and for these videos we use the whole video as our story representation

We preprocess the TRECVid collections in the same manner for all three collections The SMART stopwords[1] were removed from the ASR and the terms were stemmed using the Porter stemmer (Porter, 1980) for all documents and topic text An additional set of query stopwords,

[1]SMART stopwords, available at ftp //ftp cs cornell edu/pub/smart/

shown in Table 6, were removed from the topic text thereby producing topic text that resembles a familiar keyword style query

We will present our empirical results in the following sections, which are organised as follows Section 5 2 presents results for standard information retrieval models and for language modelling-based information retrieval models for the ASR *shot-only*, *adj-only*, *video-only* and *story-only* structures, Section 5 3 presents results using the hierarchical language models with the different hierarchical physical structures, while Section 5 4 presents results for the hierarchical language models with the different hierarchical semantic structures

5.2 Experiments with non-hierarchical structures

We can represent each shot as simply a text document consisting of either its own ASR text (*shot-only*), the enclosing video's ASR text (*video-only*), the enclosing semantic story's text (*story-only*) or the text from a sequence of shots surrounding it (*adj-only*) In this section we compare discounting-based language models (Laplace, Natural, Lidstone, Linear, Absolute), combination-based language (Witten-Bell, Jelinek-Mercer, Absolute, Dirichlet, Bayesian) and standard retrieval models (Coordinate Level Ranking, TF-IDF with log TF, BM25) for these four separate representations of the video shots, while in the following sections we will consider multi-level hierarchical representations of the video shots that combine these different representations

We summarise the average results for *shot-only*, *adj-only*, *video-only*, and *story-only* structures across the tested retrieval models in Table 7 for the TRECVid 2002, 2003 and 2004 collections The *adj-only* representation, a window of adjacent shot text, is the best physical structure and the tested retrieval models are on average 57 3%, 41 1%, and 16 8% better on this structure than on the *shot-only* structure for the TRECVid 2002, 2003 and 2004 search tasks The *video-only* structure achieves nearly the same performance as the *shot-only* structure on the TRECVid 2002 collection but it achieves very poor performance on the other collections with a MAP that is 58 6% and 88 9% lower on average than the *shot-only* structure The relatively high performance of the *video-only* structure on the TRECVid 2002 collection reflects the fact that the relevant results for some of the topics in this collection are clustered within the same videos Unsurprisingly, the semantic *story-only* structure achieves an extremely good average improvement of 85 9% compared to the *shot-only* structure for the tested retrieval models on the TRECVid 2003 collection This improvement is double the improvement achieved by using the *adj-only* structure and indicates the large potential benefit of good story segmentation to non-hierarchical text information retrieval models for the video news retrieval task

We present the results for the retrieval models for the three TRECVid collections in Figure 23 The unbiased results for these models are presented on the left side of the figure, while on the right of the figure we show the optimised results These results for individual retrieval models indicate the consistent trend of the *adj-only* structure outperforming the *shot-only* and *video-only* structures for all TRECVid collections The only exception is the Coordinate Level Ranking retrieval model on the TRECVid 2004 collection, which has better results for the *shot-only* structure than for the *adj-only* structure The Coordinate Level Ranking retrieval model as well as the MLE language model are consistently poor retrieval models for all three TRECVid collection and therefore we do not consider them as informative for deciding between

Table 7: Comparison of the average results across retrieval models for the *shot-only*, *adj-only*, *video-only* and *story-only* ASR text representations on the (a) TRECVid 2002, (b) TRECVid 2003 and (c) TRECVid 2004 collections. The percentage improvement is relative to the MAP of the *shot-only* representation

(a) TRECVid 2002 average results across retrieval models.

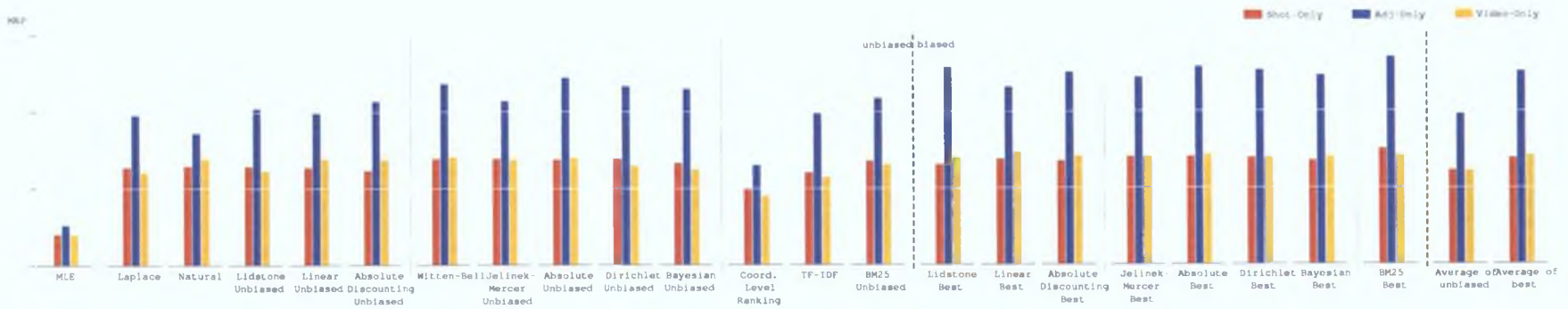
TRECVid 2002	*MAP*	*P@10*	*P@30*	*P@100*	*Impr.*
Shot-Only	**.0614**	.177	.089	.043	
Optimised	(**.0694**	.190	.100	.047)	
Adj-Only	**.979**	.174	.134	.081	+57.3%
Optimised	(**.1258**	.230	.161	.078)	+81.5%
Video-Only	**.0609**	.050	.064	.089	-0.9%
Optimised	(**.0709**	.059	.079	.101)	+2.4%

(b) TRECVid 2003 average results across retrieval models.

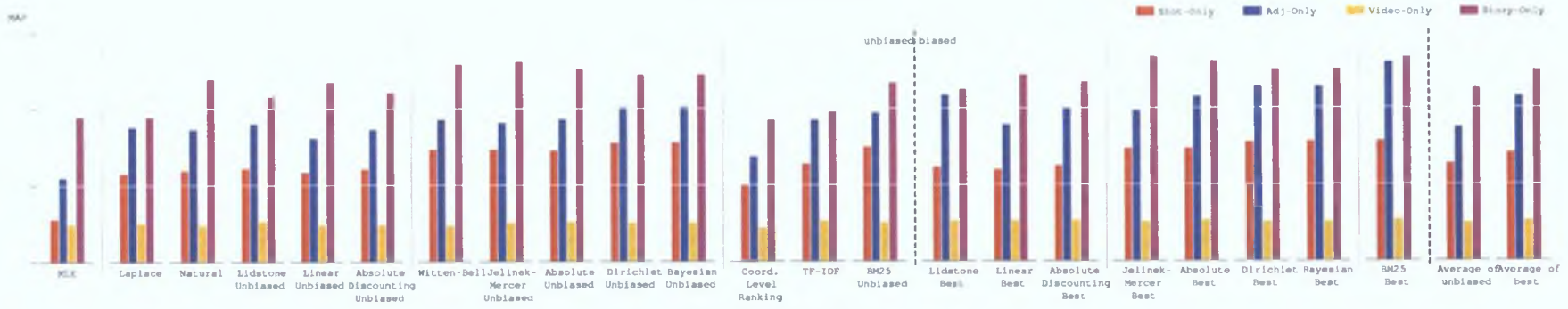
TRECVid 2003	*MAP*	*P@10*	*P@30*	*P@100*	*Impr.*
Shot-Only	**.0633**	.151	.085	.051	
Optimised	(**.0705**	.169	.098	.057)	
Adj-Only	**.0873**	.166	.115	.062	+41.1%
Optimised	(**.1077**	.157	.122	.078)	+53.2%
Video-Only	**.0247**	.033	.039	.025	-58.6%
Optimised	(**.0261**	.035	.043	.028)	-62.5%
Story-Only	**.1128**	.179	.122	.092	+85.9%
Optimised	(**.1248**	.227	.149	.105)	+78.1%

(c) TRECVid 2004 average results across retrieval models.

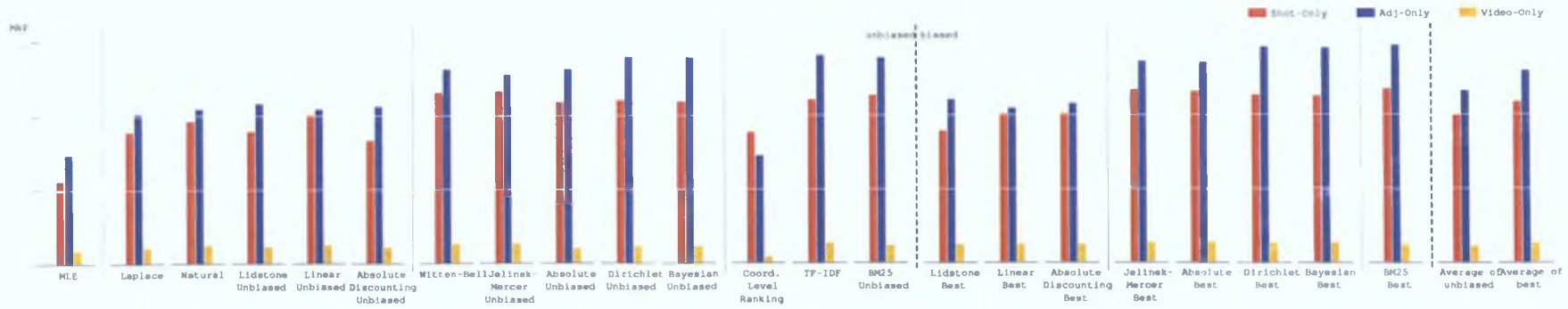
TRECVid 2004	*MAP*	*P@10*	*P@30*	*P@100*	*Impr.*
Shot-Only	**.0397**	.149	.108	.046	
Optimised	(**.0433**	.164	.116	.050)	
Adj-Only	**.0463**	.134	.094	.061	+16.8%
Optimised	(**.0516**	.133	.104	.072)	+18.9%
Video-Only	**.0044**	.020	.016	.013	-88.9%
Optimised	(**.0051**	.024	.020	.014)	-88.2%



(a) TRECVid 2002



(b) TRECVid 2003



(c) TRECVid 2004

Figure 23: Comparison of retrieval models on the ASR *shot-only*, *adj-only*, *video-only* and *story-only* structures for the (a) TRECVid 2002, (b) TRECVid 2003 and (c) TRECVid 2004 collections.

document representations The *video-only* structure is consistently poorer than the *shot-only* structure for all retrieval models on TRECVid 2003 and TRECVid 2004 collections but on the TRECVid 2002 collection it sometimes performs slightly better than the *shot-only* structure The TRECVid 2003 results indicate that when manually annotated story boundaries are available, the *story-only* structure is consistently the best representation for all retrieval models We will have to perform additional experiments in the future to see whether these results transfer to automatically detected stories, which would introduce some level of noise into the *story-only* structure

We will now investigate the retrieval models for the *adj-only* structure since it is consistently the best physical representation, while later in this section we will look more closely at the results for the semantic *story-only* structure The results for the retrieval models for the *adj-only* structure are shown in the appendices in Tables 31, 32 and 33 for TRECVid 2002, 2003 and 2004

The results for TRECVid 2002 in Table 31 (appendix page 236) show that the *adj-only* structure is statistically significantly better than the *shot-only* structure for all retrieval models except for the poorly performing Coordinate Level Ranking model The optimised results mostly use an adjacent window with a radius of 2 shots (5 shot diameter) The statistical tests comparing the unbiased retrieval models in Table 35 (appendix page 240) show that the Absolute interpolation language model is the best retrieval model with a MAP of 0 1221 and is statistically significantly better than all but the BM25 and TF-IDF retrieval models The BM25 retrieval model is the fifth best model with a MAP of 0 1088, which is 10 9% worse than the best result, while the TF-IDF model is tenth best with a MAP of 0 0988 that is 19 1% worse than the best result Both the BM25 and TF-IDF retrieval models are not statistically significantly poorer than any other retrieval model The combination-based language models are all better than the discounting based language models except for the Jelinek-Mercer language model which achieves the same MAP as the Absolute Discounting language model of 0 1069, that is 12 5% lower than the best results

The results for TRECVid 2003 in Table 32 (appendix page 237) show that all the retrieval models are statistically significantly better on the *adj-only* representation than on the *video-only* representation but all the improvements relative to the *shot-only* representation are not statistically significant This indicates that while the *adj-only* representation increases MAP by 41 1% on average relative to the *shot-only* representation on this collection, this increase is inconsistent across topics and in fact the *adj-only* representation negatively impacts a significant proportion of these topics This is the opposite pattern as observed for TRECVid 2002 and is due to the fact that the television news videos in the TRECVid 2003 collection consist of topically independent stories The *adj-only* structure therefore improves some shot's representations but when the window of adjacent shots crosses the story boundaries it adds noise to other shot's representations This table also shows that the optimised retrieval models in general use an adjacent shot window with a radius of 4 shots (9 shot diameter) on the TRECVid 2003 collection In Table 36 (appendix page 241) we compare the unbiased retrieval models on this collection The Bayesian language model achieves the best result with a MAP of 0 1010 that is statistically significantly better than all other language models except for the Absolute interpolation and Dirichlet language models The related Dirichlet language model is second best and achieves a similar MAP of 0 1008 and has the same statistical significance relationships Similar to TRECVid 2002, the

BM25 at third best and TF-IDF at sixth best are not statistically significantly poorer than any other retrieval model on this collection and achieve a MAP that is only 4 1% and 8 2% respectively worse than the Bayesian model The Absolute interpolation language model, the best model for TRECVid 2002, is fourth best for TRECVid 2003 with a MAP of 0 0929, 8 0% lower than the Bayesian model, and is statistically significantly poorer than the BM25 model All combination language models are once again better than all discounting based language models

The results for TRECVid 2004 in Table 33 (appendix page 238) show that the improvement of the retrieval models on the *adj-only* structure compared to the *shot-only* structure are not statistically significant, while the improvements relative to the *video-only* structure are all statistically significant This is completely consistent with the results for TRECVid 2003 and again indicates a large but inconsistent improvement in using the *adj-only* structure compared to the *shot-only* structure The optimum number of adjacent shots is different for this collection and in general the *adj-only* representation is best with a radius of 3 shots (diameter of 7 shots) for this collection We compare the unbiased retrieval models for this collection in Table 37 (appendix page 242) The TF-IDF model is the best retrieval model with a MAP of 0 0563, which is statistically significantly better than the Jelinek-Mercer language model and the discounting-based language models The next best retrieval models are Dirichlet (MAP 0 0559, -0 7%), Bayesian (MAP 0 0557, -1 1%), and BM25 (MAP 0 0556, -1 3%) The top four models are basically equivalent achieving similar MAP and have no statistically significant difference between them Again all combination-based language models are better than the discounting-based language models on this collection

The three collections differ in the optimum number of shots in the adjacent shot text window, which in general varies from a radius of 2 to 4 shots The unbiased results that we previously presented use the median of the best size of the optimised retrieval models' *adj-only* structures in the respective training collection We do this to reduce the noise in our comparisons of the unbiased retrieval models since otherwise differences in the number of adjacent shots for each unbiased retrieval model would account for a significant amount of the performance differences between these unbiased retrieval models For TRECVid 2002 and TRECVid 2004 the training collection is TRECVid 2003 and therefore our unbiased results for the *adj-only* structure has a radius of 4 shots on these collections, since this was the median size of the *adj-only* structure for the optimised models on training collection (i e TRECVid 2003) Likewise the unbiased models on the TRECVid 2003 collection are tuned using the TRECVid 2002 collection and therefore use an *adj-only* structure with a radius 2 shots From looking at the results for the three collections, we believe that a radius of three shots would be the best compromise for a single setting of the *adj-only* representation for future test collections

In table 8 we aggregate the topic results from the three TRECVid collections and perform statistical significance tests to compare retrieval models These aggregated TRECVid results show that the Dirichlet language model achieves the best overall result in terms of MAP In fact the Dirichlet, Bayesian, Absolute interpolation, BM25 and TF-IDF models all perform very well and are not statistically significantly poorer than any other retrieval model We can view these retrieval models as achieving equivalent effectiveness for the TRECVid search task The Jelinek-Mercer result is 9 1% lower than the best overall result and is statistically significantly poorer than the other better retrieval models (Dirichlet, Bayesian, Absolute interpolation, Witten-Bell and BM25) Lidstone is the best discounting-based approach and is followed by all other

Table 8: Statistical significance tests comparing standard retrieval models and language models using the *adj-only* structure for the aggregated TRECVid 2002, 2003 and 2004 search task. Underlined entries are significant according to the one tailed Wilcoxon Sign Rank test with a 95% confidence ($p \leq 0.05$).

Retrieval Method	*AP/%Dif*	*Wilcoxon Test Results*													
Dirichlet (Dir)		-	>Bay	>Abs	>WB	>BM25	>JM	>TF	>Lid	>AbsD	>Lap	>Lin	>Nat	>CLR	>ML
(Dir)	0.0921		.628	.186	.008	.178	.001	.198	.000	.004	.000	.000	.000	.000	.000
Bayesian (Bay)		<Dir	-	>Abs	>WB	>BM25	>JM	>TF	>Lid	>AbsD	>Lap	>Lin	>Nat	>CLR	>ML
(Bay)	0.0915/-0.7%	.628		.260	.010	.248	.002	.334	.001	.006	.000	.000	.000	.000	.000
Absolute (Abs)		<Dir	<Bay	-	>WB	>BM25	>JM	>TF	>Lid	>AbsD	>Lap	>Lin	>Nat	>CLR	>ML
(Abs)	0.0902/-2.0%	.186	.260		.269	.931	.042	.500	.000	.000	.000	.000	.000	.000	.000
Witten-Bell		<Dir	<Bay	<Abs	-	>BM25	>JM	>TF	>Lid	>AbsD	>Lap	>Lin	>Nat	>CLR	>ML
(WB)	0.0889/-3.5%	.008	.010	.269		.784	.008	.713	.007	.003	.000	.000	.000	.000	.000
BM25		<Dir	<Bay	<Abs	<WB	-	>JM	>TF	>Lid	>AbsD	>Lap	>Lin	>Nat	>CLR	>ML
(BM25)	0.0880/-4.5%	.178	.248	.931	.784		.024	.406	.000	.001	.000	.000	.000	.000	.000
Jelinek-Mercer (JM)		<Dir	<Bay	<Abs	<WB	<BM25	-	>TF	>Lid	>AbsD	>Lap	>Lin	>Nat	>CLR	>ML
(JM)	0.0837/-9.1%	.001	.002	.042	.008	.024		.890	.032	.049	.012	.000	.001	.000	.000
TF-IDF		<Dir	<Bay	<Abs	<WB	<BM25	<JM	-	>Lid	>AbsD	>Lap	>Lin	>Nat	>CLR	>ML
(TF)	0.0832/-9.6%	.198	.334	.500	.713	.406	.890		.021	.011	.000	.003	.003	.000	.000
Lidstone (Lid)		<Dir	<Bay	<Abs	<WB	<BM25	<JM	<TF	-	>AbsD	>Lap	>Lin	>Nat	>CLR	>ML
(Lid)	0.0795/-13.6%	.000	.001	.000	.007	.000	.032	.021		.277	.001	.011	.009	.000	.000
Absolute Discounting (AbsD)		<Dir	<Bay	<Abs	<WB	<BM25	<JM	<TF	<Lid	-	>Lap	>Lin	>Nat	>CLR	>ML
(AbsD)	0.0795/-13.7%	.004	.006	.000	.003	.001	.049	.011	.277		.039	.005	.007	.000	.000
Laplace		<Dir	<Bay	<Abs	<WB	<BM25	<JM	<TF	<Lid	<AbsD	-	>Lin	>Nat	>CLR	>ML
(Lap)	0.0763/-17.1%	.000	.000	.000	.000	.000	.012	.000	.001	.039		.239	.181	.003	.000
Linear (Lin)		<Dir	<Bay	<Abs	<WB	<BM25	<JM	<TF	<Lid	<AbsD	<Lap	-	>Nat	>CLR	>ML
(Lin)	0.0749/-18.7%	.000	.000	.000	.000	.000	.000	.003	.011	.005	.239		.326	.009	.000
Natural		<Dir	<Bay	<Abs	<WB	<BM25	<JM	<TF	<Lid	<AbsD	<Lap	<Lin	-	>CLR	>ML
(Nat)	0.0722/-21.6%	.000	.000	.000	.000	.000	.001	.003	.009	.007	.181	.326		.009	.000
Coord. Level Ranking		<Dir	<Bay	<Abs	<WB	<BM25	<JM	<TF	<Lid	<AbsD	<Lap	<Lin	<Nat	-	>ML
(CLR)	0.0550/-40.3%	.000	.000	.000	.000	.000	.000	.000	.000	.000	.003	.009	.009		.000
MLE		<Dir	<Bay	<Abs	<WB	<BM25	<JM	<TF	<Lid	<AbsD	<Lap	<Lin	<Nat	<CLR	-
(ML)	0.0370/-59.8%	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	

discounting models Coordinate level ranking and MLE are unsurprisingly the worst retrieval models

In summary the Dirichlet, Bayesian and Absolute language models seem to be superior to the Jelinek-Mercer language model at least for the *adj-only* representation We view these results as important in motivating our experimentation with the hierarchical versions of the Dirichlet and Absolute smoothing models m the later sections in this chapter Our experiments with the three TRECVid collections further show that none of the language models are statistically significantly better than the BM25 or the TF-IDF retrieval models and that the combination-based language models are superior to the discounting-based language models for the ASR text feature on the *adj-only* representation

We will now look at the results for the semantic *story-only* representation, which we present in Table 34 (appendix page 239) for TRECVid 2003 collection The retrieval models on the *story-only* structure are on average 30 8% better than for the physical *adj-only* structure and are statistically significantly better than all *video-only* results and nearly all unbiased *adj-only* results The retrieval models achieve on average a 85 9% improvement in MAP compared to the *shot-only* structure but similar to the *adj-only* structure on this collection, these improvements are mostly not statistically significant This is surprising given the topically cohesive nature of the story-unit, but it must be remembered that TRECVid topics are visually oriented results and that the shots within television news stories are visually very different (some are anchorperson shots, others are reports, interviews, outside broadcasts, etc) We compare the retrieval models in Table 38 (appendix page 243), which show that the Jelinek-Mercer language model achieves the highest result with a MAP of 0 1305, though this result is only statistically significantly better than the discounting-based language models The Witten-Bell language model (MAP 0 1290, -1 2%) is the second best and is followed by our previous three best language models Absolute interpolation (MAP 0 1257, -3 7%), Bayesian smoothing (MAP 0 1222, -6 4%) and Dirichlet smoothing (MAP 0 1220, -6 5%) The Bayesian language model is statistically significantly worse than the Witten-Bell language model, while the Dirichlet language models is statistically significantly worse than both Witten-Bell and the Bayesian models, though the magnitude of the difference with respect to the Bayesian model is too small to matter The BM25 model (MAP 0 1166, -10 7%) is worse at rank 8 and is only statistically significantly worse than Witten-Bell language model Again the discounting-based language models dominate the bottom half of the retrieval models The MLE language model (MAP 0 0945, -27 6%) actually achieves a higher MAP than the Laplace and Coordinate Level Ranking models though of course these differences are not significant

The unusual strong performance of the Witten-Bell language model in terms of the statistical significance tests, contradicts somewhat the ordering of retrieval models for the *adj-only* representation The differences between these representations is not due to inappropriate parameter settings even though the unbiased *story-only* retrieval models' parameters were actually set using the optimised *shot-only* representation on the TRECVid 2002 collection, since we did not have another collection with proper story-boundaries on which to tune our retrieval models The reason for this assertion is that while the optimised Absolute discounting language model achieves a higher MAP than the Witten-Bell language model and therefore shows room for improvement if it had better unbiased parameters, the Dirichlet and Bayesian language models achieve a lower optimised MAP than both the unbiased Witten-Bell and unbiased Jelinek-Mercer

language models for the *story-only* representation In other words, even with better unbiased parameters the Dirichlet and Bayesian retrieval models would still be poorer than the Witten-Bell or Jelinek-Mercer language models for the *story-only* representation

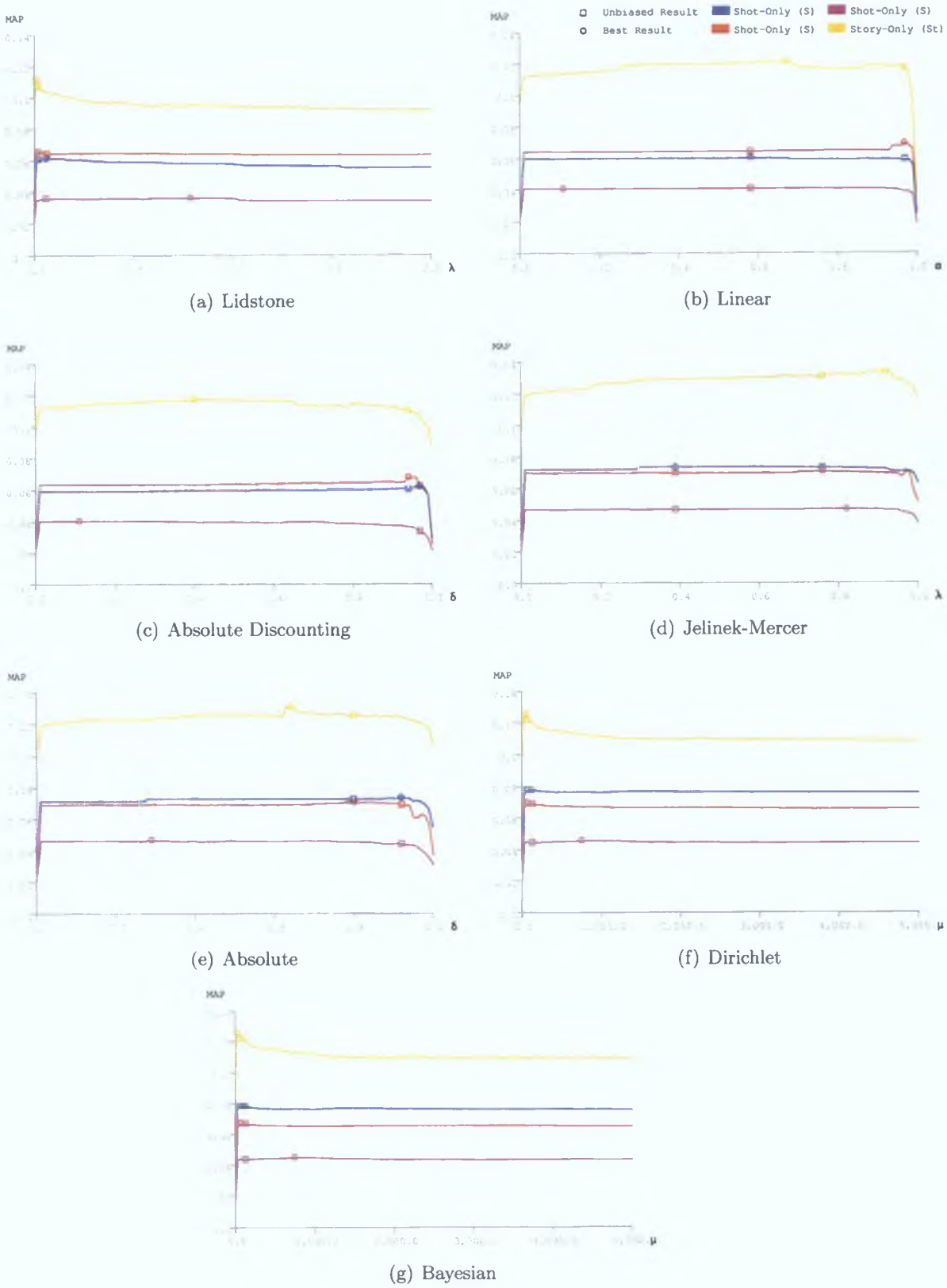
We graph the MAP response of the parametric language models to different parameter values for the *adj-only* and *story-only* representations in Figure 24, which shows that all the parametric language models have very stable performance regions for their parameters The stability of the parameters is indicated by the flatness of all the MAP curves at or around their peak values, which is particularly the case for the *adj-only* representations The optimum parameter value is sometimes dangerously near a sharp decline and we should possibly choose a safer unbiased parameter that would be further from this point and instead located within a safer more flat region of the parameter response space In this thesis we have simply chosen the optimum parameters in terms of MAP in the training collections as our unbiased parameters but it would be interesting in future to investigate possibly better methods for selecting the unbiased parameter value

We present the topic results for the Dirichlet language model on the *shot-only*, *adj-only*, *video-only* and *story-only* representations for the three TRECVid search tasks in Figure 25 These figures show that the Dirichlet language model performs better using the *shot-only* representation than for any of the other representations for 6 topics in TRECVid 2002, 8 topics in TRECVid 2003 and 14 topics in the TRECVid 2004 results These figures also show that the *adj-only* representation improves many of the topic results on all three collections compared to the *shot-only* and *video-only* representations The inconsistency where the *adj-only* representation helps some topics while hindering others was picked up previously in the failed statistical significance tests So while the average performance in terms of MAP is increased by a large amount by using the *adj-only* representation, it is still somewhat hit-or-miss for certain topics where the *shot-only* representation may actually perform better We can see from looking at the results for TRECVid 2003 that the *story-only* representation is best for only 10 topics, *shot-only* is best for 8, *adj-only* is best for 6 and *video-only* is best for a single topic Obviously if almost one third of the results are better using the *shot-only* representation we cannot consider the *story-only* representation a reliable improvement over the *shot-only* representation We hope to remedy this consistency problem by hierarchically combining representations in the following sections

5.3 Experiments with hierarchical physical video structures

In this section we will evaluate two hierarchical physical structures *shot+video* and *shot+adj+video* for the hierarchical Jelinek-Mercer language model (Westerveld, de Vries and van Ballegooij, 2003) and our proposed hierarchical Dirichlet, hierarchical Absolute and hierarchical Witten-Bell language models, which we described in the previous chapter A linear interpolation of the BM25 or TF-IDF scores for the *shot*, *adj* and *video* structures would likely achieve similar performance but we have not investigated such an approach – instead we concentrate solely on hierarchical language models in this section

The results for these four language models on the *shot+video* and *shot+adj+video* hierarchical representations for the three TRECVid collections are displayed in Figure 26 This figure



(a) Lidstone

(b) Linear

(c) Absolute Discounting

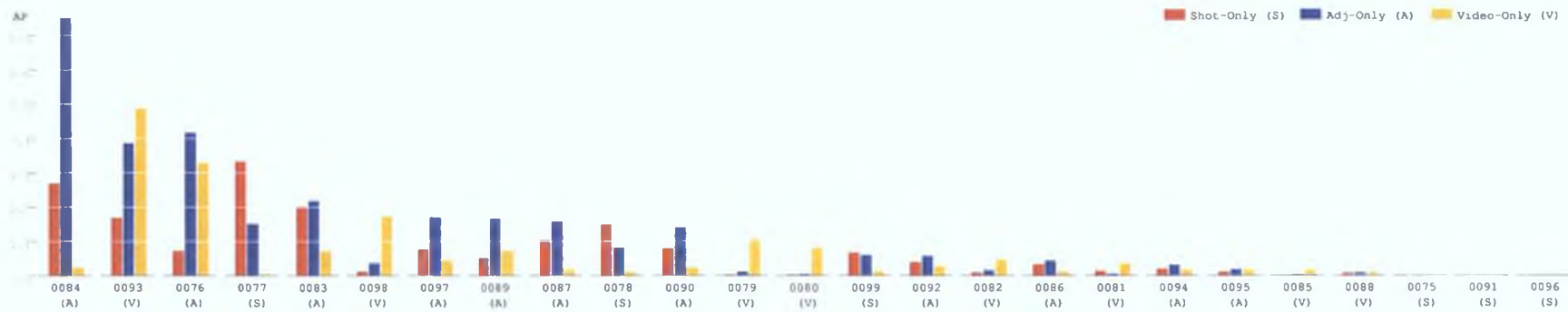
(d) Jelinek-Mercer

(e) Absolute

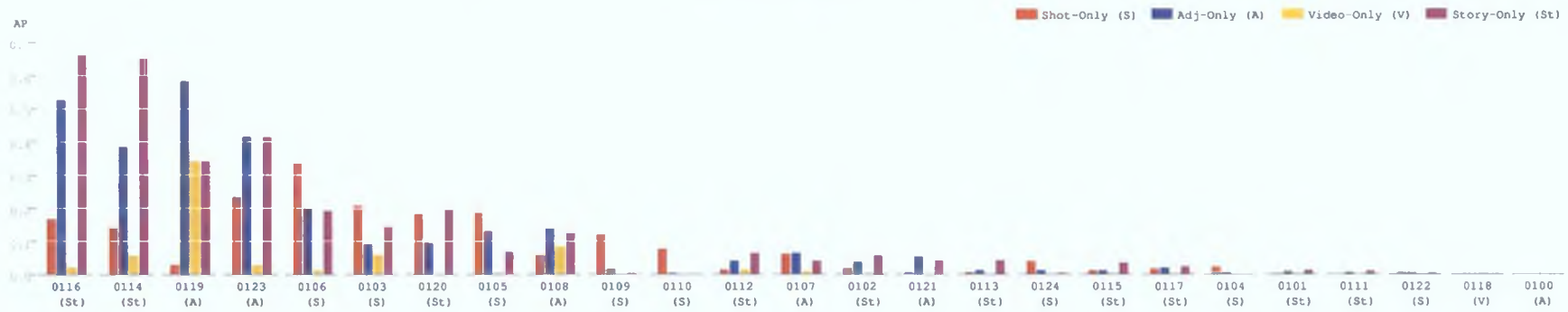
(f) Dirichlet

(g) Bayesian

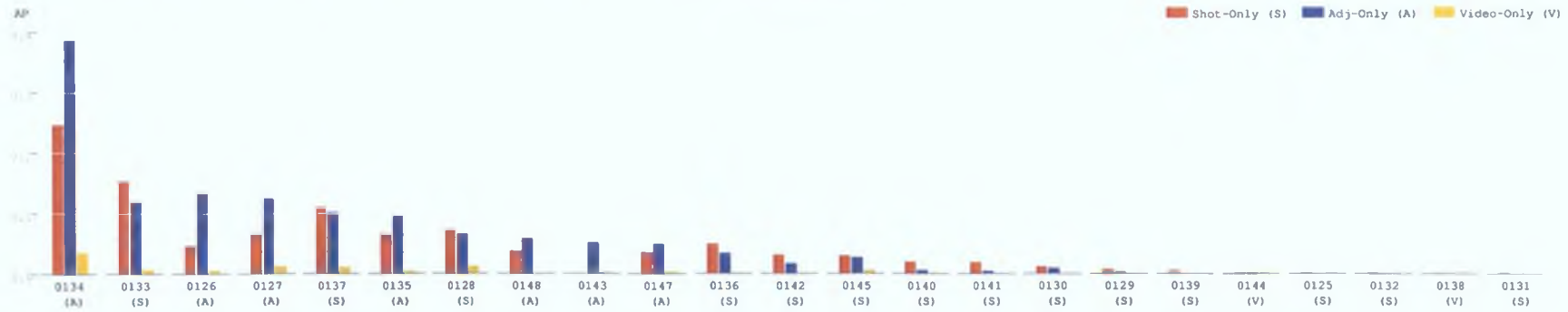
Figure 24: Response of MAP to the different parameter values for the parametric language models on the *adj-only* structure for the TRECVid 2002, 2003 and 2004 collections and for the *story-only* structure on TRECVid 2003 collection.



(a) TRECVid 2002

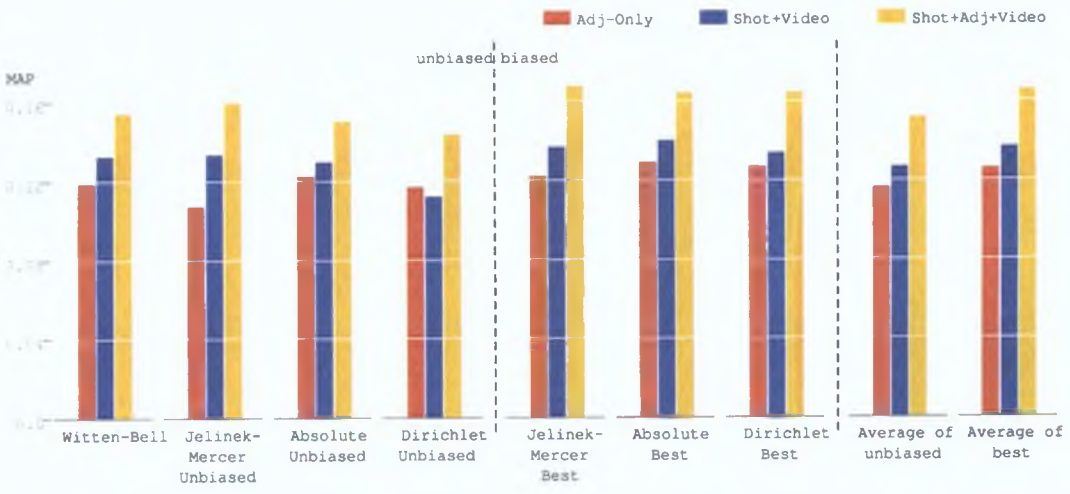


(b) TRECVid 2003

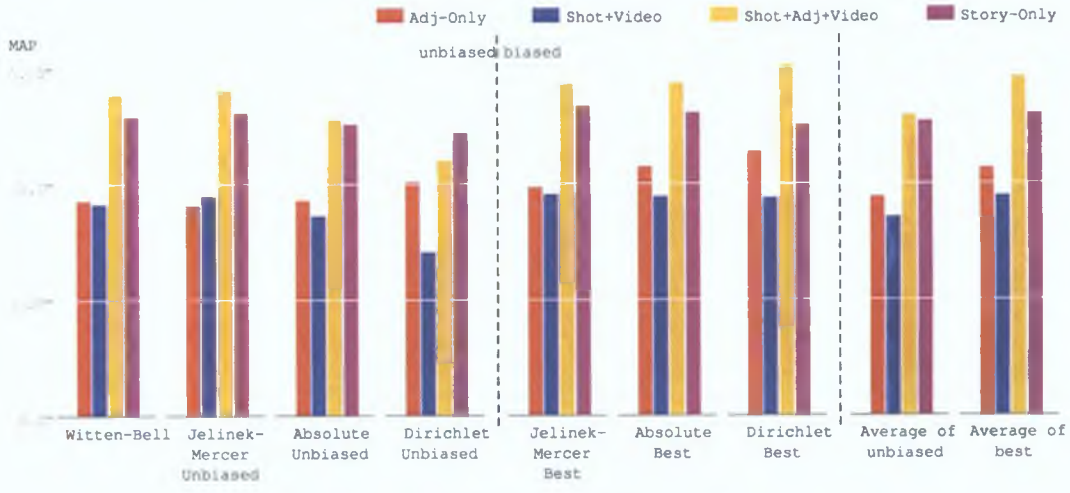


(c) TRECVid 2004

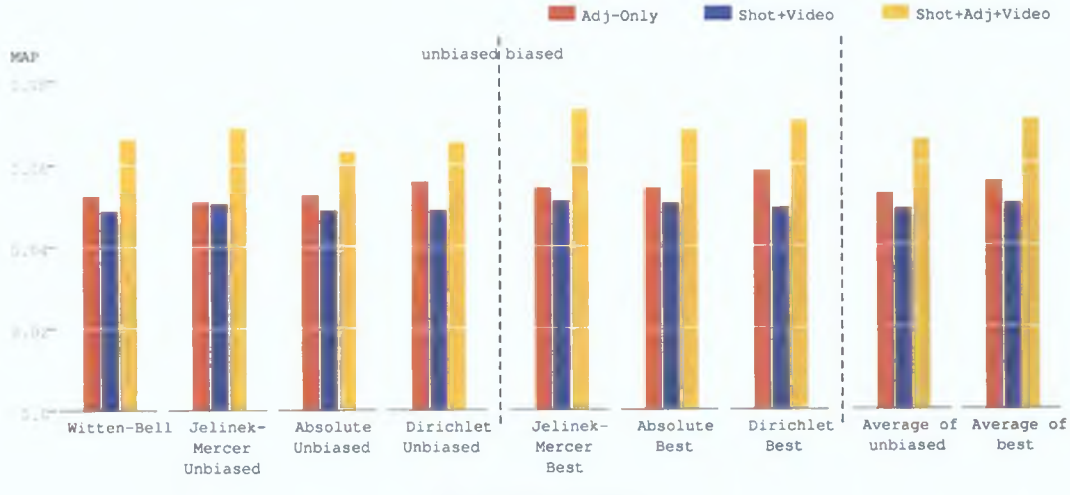
Figure 25: Comparison of the *shot-only*, *adj-only*, *video-only* and *story-only* ASR text representations for each topic using the Dirichlet language model for the (a) TRECVid 2002, (b) TRECVid 2003 and (c) TRECVid 2004 collections.



(a) TRECVid 2002



(b) TRECVid 2003



(c) TRECVid 2004

Figure 26: Comparison of structural smoothing using the *shot+video* and *shot+adj+video* hierarchical physical structures with the *adj-only* structure for the (a) TRECVid 2002, (b) TRECVid 2003 and (c) TRECVid 2004 collections

shows that the *shot+adj+video* representation consistently outperforms the *shot+video* and the *adj-only* representations for all retrieval models on the three collections and that the *shot+video* representation actually performs worse than the *adj-only* representation on the TRECVid 2003 and TRECVid 2004 collections, which is due to the multi-topic nature of television news programmes Interestingly, the results for TRECVid 2003 show that the *shot+adj+video* structure actually performs better than the semantic *story-only* representation for all retrieval models except the unbiased hierarchical Dirichlet language model We can also see that the hierarchical Jelinek-Mercer language model produces slightly better unbiased results than the other three hierarchical language models on the three collections Only the optimised results of the hierarchical Dirichlet and hierarchical Absolute language models outperform it on the TRECVid 2003 collection

We look more closely at the successful *shot+adj+video* hierarchical physical representation in Table 9 for the three TRECVid collections Remarkably the retrieval models for the *shot+adj+video* hierarchical structure are statistically significantly better than the *shot-only*, *adj-only* and *shot+video* structures for all three collections and for all retrieval models except the unbiased hierarchical Dirichlet language model on TRECVid 2003 collection This is significant as it implies that the *shot+adj+video* structure gives consistent improvement over the *shot-only* representation, which both the well performing *adj-only* and *story-only* structures failed to achieve On the TRECVid 2003 collection, the hierarchical Jelinek-Mercer language model is actually 7 6% better than the semantic *story-only* representation but this improvement is not statistically significant The improvements for the hierarchical Jelinek-Mercer language model on the *shot+adj+video* structure compared to the previously best physical non-hierarchical representation *adj-only* are quite large and statistically significant at 50 2% for TRECVid 2002, 48 6% for TRECVid 2003 and 34 4% for TRECVid 2004 Our other proposed hierarchical language models are likewise statistically significantly improved using the *shot+adj+video* representation but by not as much as the Jelinek-Mercer hierarchical language model

We investigate the individual topic results for the hierarchical Jelinek-Mercer language model on the three TRECVid collections in Figure 27 The topic results for TRECVid 2002 indicate the fairly consistent improvement of the *shot+adj+video* representation over the other physical representations It is the best representation for 13 of the 25 topics and the other 12 topics are best 3 times for *shot-only*, 2 times for *adj-only* and 7 times for *shot+video* Importantly, when the *shot+adj+video* representation is poorer than the other representations the performance difference in terms of average precision is small The topic results for TRECVid 2003 show that the *shot+adj+video* representation is best for 17 of the 25 topics and that similar to the TRECVid 2002 collection when it is less than the other representations it is by a small amount in terms of average precision The situation is similar for TRECVid 2004 in which the *shot+adj+video* representation is best for 12 of the 23 The *shot-only* representation is best for 5 of the 6 worst topics, while the *shot+video* representation is best for 2 poorly performing topics and the *adj-only* representation is better for 4 medium level topics, but only one of these topics, topic 143, shows any notable difference in performance in terms of average precision Overall, we consider the *shot+adj+video* representation a reliable improvement on the other physical representations for the individual topic results

We compare the hierarchical language models for their unbiased results on the *shot+adj+video* representation in Table 40 (appendix page 245) The hierarchical Jelinek-Mercer language model

Table 9: Comparison of the *shot+adj+vid* ASR text representation with the *shot-only*, *adj-only*, *story-only* and *shot+video* representations for the hierarchical language models on the (a) TRECVid 2002, (b) TRECVid 2003 and (c) TRECVid 2004 collections.

(a) TRECVid 2002

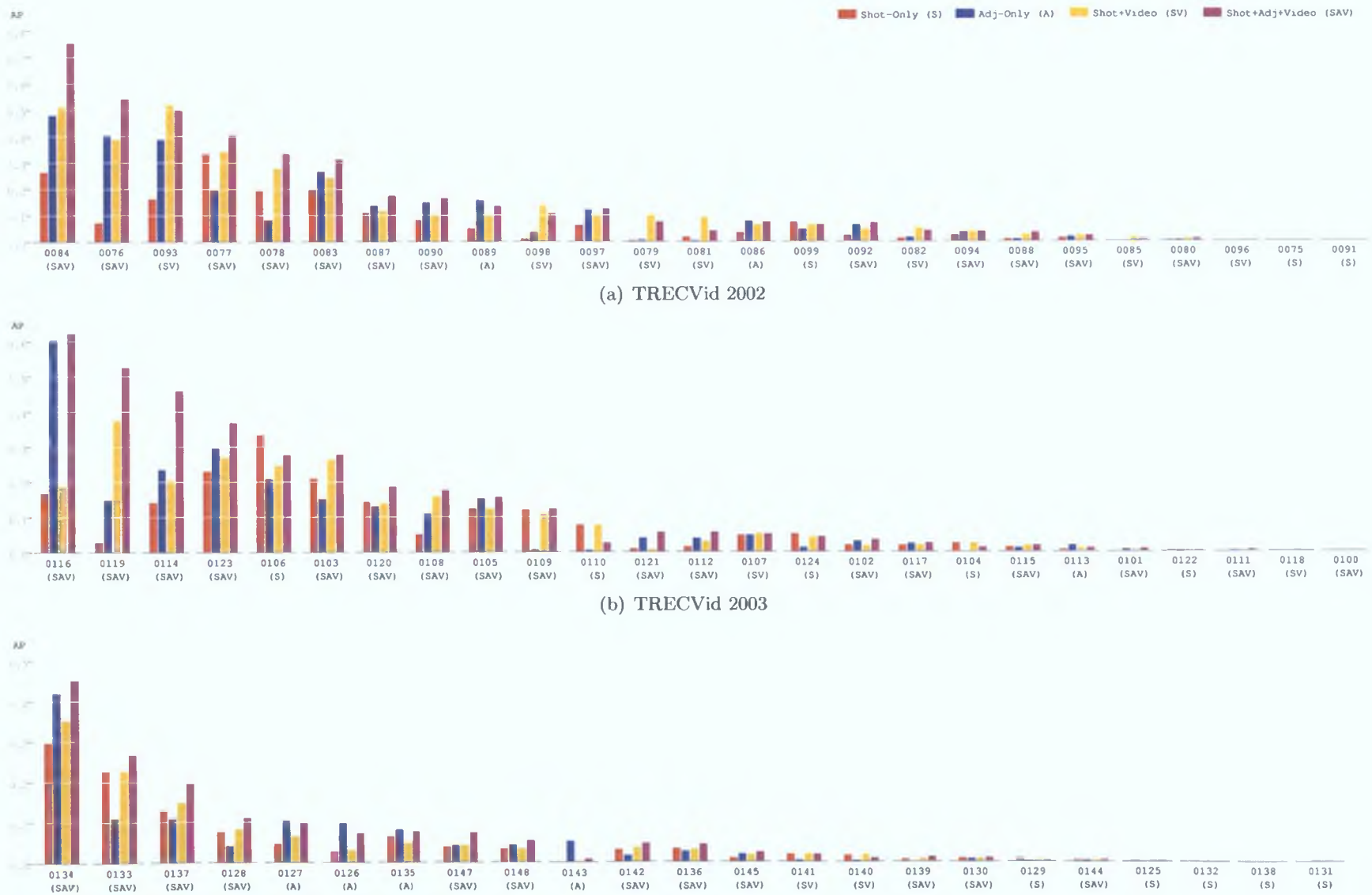
TRECVID 2002	*Shot+Adj+Video*					*V. Shot-Only*	*V. Adj-Only*	*V. Shot+Video*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*
Witten-Bell (Adj)	4	**.1541**	.260	.167	.118	+121.8% ∼ .000	+30.2% ∼ .015	+16.3% ∼ .047
Jelinek-Mercer (Adj, $\lambda_{shot}, \lambda_{adj}, \lambda_{vid}, \lambda_{col}$)	4, 0.30, 0.10, 0.10, 0.50	**.1605**	.264	.175	.117	+131.1% ∼ .000	+50.2% ∼ .000	+20.3% ∼ .015
Best	(4, 0.15, 0.15, 0.20, 0.50	**.1677**	.272	.175	.118)	+138.3% ∼ .000	+37.4% ∼ .005	+22.5% ∼ .011
Absolute (Adj, $\delta_{adj}, \delta_{vid}, \delta_{col}$)	10, 0.30, 0.80, 0.95	**.1502**	.268	.188	.126	+118.3% ∼ .000	+23.0% ∼ .030	+15.9% ∼ .007
Best	(2, 0.80, 0.90, 0.95	**.1643**	.244	.192	.122)	+133.4% ∼ .000	+27.3% ∼ .018	+17.1% ∼ .003
Dirichlet (Adj, $\mu_{adj}, \mu_{vid}, \mu_{col}$)	5, 150, 3000, 30000	**.1433**	.228	.171	.106	+107.9% ∼ .000	+22.9% ∼ .001	+28.0% ∼ .000
Best	(2, 2000, 150, 2000	**.1645**	.252	.185	.112)	+136.2% ∼ .000	+29.8% ∼ .026	+23.0% ∼ .125

(b) TRECVid 2002

TRECVID 2003	*Shot+Adj+Video*					*V. Shot-Only*	*V. Adj-Only*	*V. Shot+Video*	*V. Story-Only*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*
Witten-Bell (Adj)	3	**.1384**	.252	.176	.113	+89.3% ∼ .006	+49.3% ∼ .001	+51.6% ∼ .001	+7.3% ∼ .366
Jelinek-Mercer	**Adj, $\lambda_{shot}, \lambda_{adj}, \lambda_{vid}, \lambda_{col}$**								
Unbiased	4, 0.15, 0.15, 0.20, 0.50	**.1405**	.252	.176	.113	+92.4% ∼ .004	+55.1% ∼ .000	+48.6% ∼ .001	+7.6% ∼ .314
Best	(4, 0.30, 0.10, 0.10, 0.50	**.1429**	.240	.179	.115)	+94.2% ∼ .001	+45.2% ∼ .000	+49.7% ∼ .000	+7.2% ∼ .246
Absolute	**Adj, $\delta_{adj}, \delta_{vid}, \delta_{col}$**								
Unbiased	2, 0.80, 0.90, 0.95	**.1277**	.264	.169	.102	+76.0% ∼ .015	+37.4% ∼ .015	+48.1% ∼ .001	+1.6% ∼ .666
Best	(10, 0.30, 0.80, 0.95	**.1435**	.256	.196	.105)	+95.3% ∼ .002	+33.7% ∼ .019	+51.7% ∼ .000	+9.8% ∼ .238
Dirichlet	**Adj, $\mu_{adj}, \mu_{vid}, \mu_{col}$**								
Unbiased	2, 2000, 150, 2000	**.1100**	.208	.167	.097	+42.0% ∼ .443	+9.1% ∼ .622	+56.0% ∼ .002	-9.9% ∼ .049
Best	(5, 150, 3000, 30000	**.1510**	.260	.189	.111)	+94.5% ∼ .001	+32.8% ∼ .001	+60.8% ∼ .001	+20.6% ∼ .071

(c) TRECVid 2002

TRECVID 2004	*Shot+Adj+Video*					*V. Shot-Only*	*V. Adj-Only*	*V. Shot+Video*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*
Witten-Bell (Adj)	4	**.0660**	.200	.141	.081	+42.8% ∼ .020	+25.7% ∼ .026	+36.1% ∼ .000
Jelinek-Mercer (Adj, $\lambda_{shot}, \lambda_{adj}, \lambda_{vid}, \lambda_{col}$)	4, 0.30, 0.10, 0.10, 0.50	**.0686**	.209	.143	.091	+47.2% ∼ .001	+34.4% ∼ .006	+35.7% ∼ .000
Best	(3, 0.15, 0.15, 0.02, 0.68	**.0730**	.222	.141	.093)	+55.9% ∼ .001	+33.9% ∼ .001	+42.4% ∼ .001
Absolute (Adj, $\delta_{adj}, \delta_{vid}, \delta_{col}$)	10, 0.30, 0.80, 0.95	**.0629**	.209	.136	.083	+44.0% ∼ .003	+19.6% ∼ .026	+29.0% ∼ .000
Best	(3, 0.10, 0.10, 0.80	**.0680**	.204	.143	.087)	+46.8% ∼ .002	+25.1% ∼ .005	+34.4% ∼ .004
Dirichlet (Adj, $\mu_{adj}, \mu_{vid}, \mu_{col}$)	5, 150, 3000, 30000	**.0651**	.191	.148	.090	+47.7% ∼ .002	+16.5% ∼ .014	+33.3% ∼ .001
Best	(3, 400, 1000, 30000	**.0702**	.178	.159	.095)	+55.1% ∼ .004	+20.5% ∼ .006	+41.9% ∼ .001



(a) TRECVid 2002

(b) TRECVid 2003

(c) TRECVid 2004

Figure 27: Comparison of structural smoothing using the *shot+video* and *shot+adj+video* physical structures with the *shot-only* and *adj-only* smoothed representation for each topic using the Jelinek-Mercer (linear interpolation) language model for the (a) TRECVid 2002, (b) TRECVid 2003 and (c) TRECVid 2004 collections

performs best for the three TRECVid collections and is significantly better than the hierarchical Absolute and hierarchical Dirichlet language models on two of these collections It is also significantly better than the hierarchical Witten-Bell language model on one of the collections The hierarchical Witten-Bell language model is the second best model and is on average 3 4% worse in terms of MAP than the hierarchical Jelinek-Mercer language model, while the hierarchical Absolute language model is the third best model with on average 7 9% lower MAP than the hierarchical Jelinek-Mercer result The hierarchical Dirichlet language model performs very poorly on the TRECVid 2003 collection achieving a 21 7% poorer MAP than the hierarchical Jelinek-Mercer results and is also statistically significantly poorer than all other retrieval models In contrast on the TRECVid 2004 collection it achieves a far better MAP that is only 5 0% worse than the hierarchical Jelinek-Mercer results and that is not statistically significantly poorer than any other retrieval model This contradiction in the performance of the hierarchical Dirichlet language model on these two television news collections is due to the inappropriate unbiased parameters for this model on the TRECVid 2003 collection Even though it has the best optimised results on this collection, the hierarchical Dirichlet language model has the poorest unbiased results due to the inappropriate (and possibly over-fitted) optimised parameters from the TRECVid 2002 collection (see Table 9)

We compare the results for the hierarchical language models on the *shot+video* representation in Table 39 (appendix page 244) These results corroborate the observations for the *shot+adj+video* representation and show that the hierarchical Jelinek-Mercer language model is again the best model and is followed by on average the Witten-Bell, Absolute and Dirichlet hierarchical language models The unbiased results for the Dirichlet language model again perform poorly on the TRECVid 2003 collection and also on the TRECVid 2002 collection indicating instability in the optimum parameter settings between these two collections

We clearly see from these sets of results that the hierarchical Jelinek-Mercer language model is better than our proposed hierarchical language model variations for the physical hierarchical representations *shot+video* and *shot+adj+video* Though the differences are not very large for the hierarchical Witten-Bell and hierarchical Absolute models, we find these results quite discouraging as they are consistent in indicating that the hierarchical Jelinek-Mercer language model produces better unbiased results This is in contrast to the results for the non-hierarchical versions of these smoothing methods on the *adj-only* representation for the same three TRECVid collections, which showed superior results for these smoothing models over the Jelinek-Mercer language model

5.4 *Experiments with hierarchical semantic video structure*

In this section we compare the hierarchical language models for the semantic *shot+story* and *shot+adj+story* structures These experiments involved only the TRECVid 2003 collection as this was the only collection which had associated story boundaries The shots that belonged to more than one story were split into multiple sub-shots and the retrieval results were post-processed to remove duplicate shots by keeping the highest score for each shot The new sub-shots each contained only the segment of the ASR text from the original shot that overlapped with their respective story This allows us to have a strict hierarchy for shots (or sub-shots when we

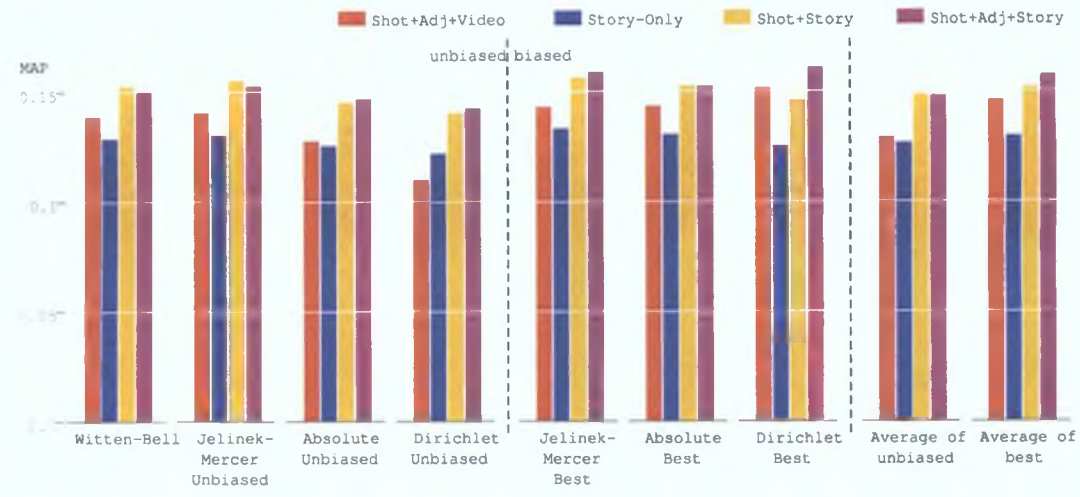


Figure 28: Comparison of hierarchical LMs on the ASR *shot+story*, *shot+adj+story* hierarchical semantic structures with the *story-only* and *shot+adj+video* structures for TRECVid 2003.

split the shot because it spans two stories) and stories in which shots (or sub-shots) only have one parent even when the original shot is really a member of multiple stories.

We compare the hierarchical language models for the semantic *shot+story* and *shot+adj+story* representations in Figure 28, which shows that the two semantic representations achieve similar results. The mean results for the semantic hierarchical representations show moderate improvements over the best physical hierarchical representation *shot+adj+video* and the semantic non-hierarchical representation *story-only*. The unbiased results indicate that the hierarchical Jelinek-Mercer language model is just slightly better than the hierarchical Witten-Bell language model and is more significantly better than the other two hierarchical smoothing models, hierarchical Absolute and hierarchical Dirichlet. The optimised results show that the hierarchical Dirichlet language model can achieve a potentially higher MAP than the hierarchical Jelinek-Mercer language model for the *shot+adj+story* representation.

In Table 10 we present the results for the *shot+story* and *shot+adj+story* hierarchical structures. The hierarchical Jelinek-Mercer language model achieves a MAP of 0.1551 on the *shot+story* structure, which is statistically significantly better than its results on the *shot+adj+story* and *shot+adj+video* representations by 1.6% and 10.4% respectively. The 10.4% improvement on the physical structure *shot+adj+video* is perhaps disappointing when we consider that the *shot+story* representation uses the correctly identified story boundaries. Interestingly the inclusion of the *adj* estimator in the semantic structure does not improve the results and in fact decreases the results for some of the hierarchical retrieval models such as the Witten-Bell and Jelinek-Mercer language models. All retrieval models are statistically significantly better on the *shot+story* and *shot+adj+story* representations than on the *shot-only* representation. In addition all unbiased retrieval models except hierarchical Absolute for the *shot+story* representation are statistically significantly better on these representations than on the best hierarchical physical representation *shot+adj+video*. So while the improvements in the retrieval models on this collection are about a modest 10%, these improvements reflect a consistent performance boost across the individual topics in the TRECVid 2003 collection. The results show that there is not much difference between the *shot+story* and *shot+adj+story* representations for the hierarchical Witten-Bell, hierarchical Jelinek-Mercer and hierarchical Absolute language models.

Table 10: Comparison of the (a) *shot+story* and (b) *shot+adj+story* ASR text representation with each other and with the *shot-only* and *story-only* representations for the hierarchical language models on the TRECVid 2002 collection.

(a) Shot+Story

TRECVID 2003	*Shot+Story*					*V. Shot-Only*	*V. Shot+Adj+Video*	*V. Story-Only*	*V. Shot+Adj+Story*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*
Witten-Bell		**.1526**	.248	.211	.124	+108.7% ∼ .001	+10.2% ∼ .015	+18.3% ∼ .003	+1.8% ∼ .377
Jelinek-Mercer	$\lambda_{shot}, \lambda_{story}, \lambda_{col}$								
Unbiased	0.17, 0.23, 0.60	**.1551**	.264	.213	.122	+112.4% ∼ .000	+10.4% ∼ .043	+18.8% ∼ .003	+1.6% ∼ .024
Best	(0.08, 0.16, 0.76	**.1560**	.264	.207	.122)	+112.1% ∼ .001	+9.2% ∼ .014	+17.0% ∼ .001	-1.5% ∼ .977
Absolute	$\delta_{story}, \delta_{col}$								
Unbiased	0.90, 0.87	**.1453**	.232	.185	.119	+100.2% ∼ .006	+13.8% ∼ .052	+15.6% ∼ .011	-1.0% ∼ .545
Best	(0.30, 0.40	**.1523**	.248	.204	.126)	+107.3% ∼ .001	+6.1% ∼ .116	+16.6% ∼ .049	+0.2% ∼ .036
Dirichlet	μ_{story}, μ_{col}								
Unbiased	350, 1750	**.1403**	.208	.195	.118	+81.1% ∼ .011	+27.5% ∼ .000	+14.9% ∼ .046	-1.5% ∼ .736
Best	(50, 3250	**.1463**	.232	.199	.115)	+88.4% ∼ .000	-3.1% ∼ .642	+16.8% ∼ .170	-8.9% ∼ .041

(b) Shot+Adj+Story

TRECVID 2003	*Shot+Adj+Story*					*V. Shot-Only*	*V. Shot+Adj+Video*	*V. Story-Only*	*V. Shot+Story*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*
Witten-Bell (Adj)	3	**.1499**	.244	.196	.125	+105.0% ∼ .003	+8.3% ∼ .010	+16.2% ∼ .010	-1.8% ∼ .377
Jelinek-Mercer	**Adj**, $\lambda_{shot}, \lambda_{adj}, \lambda_{story}, \lambda_{col}$								
Unbiased	4, 0.15, 0.15, 0.20, 0.50	**.1526**	.256	.192	.128	+109.0% ∼ .002	+8.6% ∼ .023	+16.9% ∼ .014	-1.6% ∼ .024
Best	(15, 0.10, 0.15, 0.02, 0.72	**.1584**	.272	.205	.116)	+115.3% ∼ .000	+10.8% ∼ .024	+18.8% ∼ .002	+1.5% ∼ .977
Absolute	**Adj**, $\delta_{adj}, \delta_{story}, \delta_{col}$								
Unbiased	2, 0.80, 0.90, 0.95	**.1467**	.224	.196	.124	+102.2% ∼ .006	+14.9% ∼ .006	+16.8% ∼ .021	+1.0% ∼ .545
Best	(15, 0.10, 0.90, 0.30	**.1520**	.244	.201	.115)	+106.9% ∼ .001	+5.9% ∼ .115	+16.4% ∼ .081	-0.2% ∼ .036
Dirichlet	**Adj**, $\mu_{adj}, \mu_{story}, \mu_{col}$								
Unbiased	2, 2000, 150, 2000	**.1424**	.224	.187	.129	+83.9% ∼ .032	+29.5% ∼ .000	+16.7% ∼ .017	+1.5% ∼ .736
Best	(5, 250, 2000, 7500	**.1606**	.240	.203	.128)	+106.8% ∼ .001	+6.3% ∼ .019	+28.3% ∼ .014	+9.8% ∼ .041

We compare the hierarchical retrieval models for the *shot+story* representation in Table 41 (appendix page 246) and for the *shot+adj+story* representation in Table 42 (appendix page 246) The results for the retrieval models for both representations are a lot closer than previously presented for the hierarchical physical representations The best result for the *shot+video* representation is hierarchical Jelinek-Mercer (MAP 0 1551), the second best is hierarchical Witten-Bell (MAP 0 1526, -1 6%), which is followed by hierarchical Absolute (MAP 0 1453, -6 3%) and the worst result is hierarchical Dirichlet (MAP 0 1403, -9 6%) The order of retrieval models is exactly the same for the *shot+adj+video* representation For both hierarchical semantic representations the only retrieval model to be statistically significantly worse than another is the hierarchical Dirichlet language model, which is statistically significantly worse than the Jelinek-Mercer and Witten-Bell language models for the *shot+video* representation and is statistically significantly worse than the hierarchical Jehnek-Mercer language for the *shot+adj+video* representation These results show that the hierarchical Jelinek-Mercer, Witten-Bell and Absolute language models achieve somewhat equivalent results for the hierarchical semantic structures *shot+story* and *shot+adj+story*

We compare the topic results of the Jelinek-Mercer language model for the *shot+story*, *shot+adj+story*, *shot+adj+video* and the *story-only* representations in Figure 29 The *shot+story* representation is best for 11 of the 25 topics and on only one topic, topic 119, is another representation notably better The topic 119 for shots with Morgan Freeman actually has all its 18 relevant shots in a single C-SPAN programme that lacked story boundaries and therefore this topic is not indicative of the relative performance of the different semantic and physical representations These topic results show that the *shot+story* representation achieves a very stable performance increase relative to the other representations

In summary, the *shot+story* structure is the best performing hierarchical semantic representation and the retrieval models using it are mostly statistically significantly better than for all other representations The difference between this representation and the best hierarchical physical representation *shot+adj+video* is a statistically significant 10 4% for the Jelinek-Mercer language model, which is perhaps disappointing considering we are using the correct story boundaries Viewed from another perspective it is a positive endorsement of the *shot+adj+video* representation that it can achieve such relatively high results without any semantic information

The hierarchical Jehnek-Mercer language model is the best performing retrieval model for these hierarchical semantic structures The differences between the different retrieval models are a lot smaller for this semantic representation than for the hierarchical physical representations and in fact the hierarchical Dirichlet model is the only statistically significantly poorer retrieval model for the semantic hierarchical representations Our proposed hierarchical Witten-Bell and hierarchical Absolute language models achieve only slightly poorer results than the hierarchical Jelinek-Mercer language model and these retrieval models could possibly be considered equivalent in terms of performance

It may be the case that a semantic hierarchy is more conducive to different hierarchical smoothing functions and that the more topically noisy *shot+adj+video* representation is only suitable for the Jelmek-Mercer language model, a linear interpolation of the probability estimators that keeps the effects of any topical noise from the different hierarchy levels, *adj* and *video*, constant for each retrieval document's representation The other hierarchical smoothing

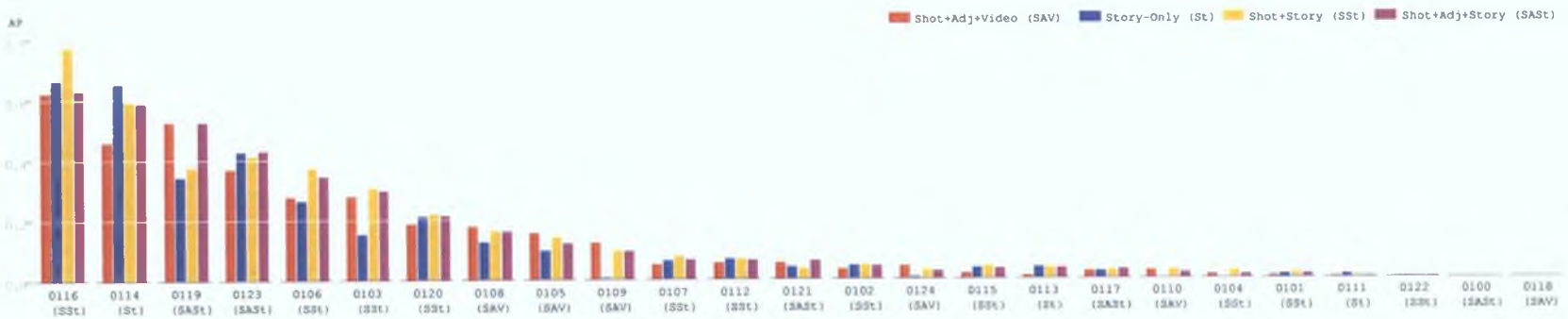


Figure 29: Comparison of Jelinek-Mercer LM topic results for the ASR *story-only*, *shot+story*, *shot+adj+story* and *shot+adj+video* structures on TRECVid 2003.

methods may be ineffective for non-semantic hierarchies as they vary the amounts of smoothing depending on the distribution of text at each level in the hierarchy and therefore the effects of topical noise caused by the *adj* or *video* representations crossing multiple story boundaries will be further increased with the effects of the changing amounts of smoothing at each level We believe that a single set of experiments on the TRECVid 2003 collection is an unstable basis on which to build an elaborate explanation and that further studies of the *shot+story* and *shot+adj+shory* representations would be informative With that caveat noted, we believe that the different smoothing methods are probably more suitable for the semantic hierarchical structures than for the physical hierarchical structures

5.5 Summary

The best non-hierarchical retrieval structure for the video shot retrieval task is to represent each shot using a window of adjacent shot text This *adj-only* structure achieves on average a 57 3%, 41 1% and 16 8% improvement on using the shot text alone for the three TRECVid collections but these improvements are not statistically significant for TRECVid 2003 and 2004 collections The Dirichlet language model was overall the best retrieval model for the *adj-only* representation for the TRECVid search tasks, though the Dirichlet, Bayesian, Absolute interpolation, BM25 and TF-IDF models all perform very well and are not statistically significantly poorer than any other retrieval model The Jelinek-Mercer language model was slightly poorer than most of these models The discounting-based language models performed worse than these retrieval models and Coordinate Level Ranking and MLE were the worst performing retrieval models We view these results as vindication of the benefits of applying the language modelling approach to video retrieval as well as an indication that language models other than Jelinek-Mercer have a potential benefit in the video shot retrieval task

The semantic story unit provides an even better performance increase than for any of the physical non-hierarchical representations It improves on the shot representation on average by 85 9% and improves on the *adj-only* representation by 30 8% For most retrieval models the improvement relative to the *adj-only* representation are statistically significant whereas importantly the improvement relative to the *shot-only* representation are not in general statistically significant This indicates that while the *story-only* representation improves the average performance for the video retrieval tasks, it does not do so reliably across the TRECVid topics We found that the Jehnek-Mercer language model produces the best results for the *story-only* representation The other combination-based language models Witten-Bell, Absolute interpolation, Bayesian and Dirichlet achieve the next best results and are not statistically significantly poorer than Jelinek-Mercer The BM25 algorithm is eight best and was only statistically significantly poorer than the Witten-Bell language model, while TF-IDF is the eleventh best retrieval model and was statistically significantly poorer than most of the higher performing retrieval models The combination-based language models achieve better performance than all the discounting-based language models

The best hierarchical physical structure for video shot retrieval is the *shot+adj+video* The hierarchical Jelinek-Mercer language model produces statistically significantly better results for this representation than for the *shot-only*, *adj-only* and *shot+video* physical structures on all

three collections It even achieves an improvement of 7 6% over the semantic *story-only* representation, though this is not statistically significant

The best hierarchical semantic structure for video shot retrieval is the *shot+story* representation, which achieves statistically significantly better results than the *shot-only* (+112 4), *shot+adj+video* (+10 4%), *story-only* (+18 8) and *shot+adj+story* (+1 6%) representations for the hierarchical Jelinek-Mercer language model on the TRECVid 2003 collection

We found that our hierarchical language models did not perform as well as the hierarchical Jelinek-Mercer (linear interpolation) language model on the physical and semantic hierarchical structures The results for the semantic hierarchies were a lot closer than the physical hierarchies and indicate a potential equivalency in performance terms between the hierarchical Jelinek-Mercer, hierarchical Witten-Bell and hierarchical Absolute language models It may be the case that Jelinek-Mercer is the best language model for the physical hierarchy due to the fact that it keeps the influence of each level in the hierarchy constant between the document structures While the *adj* and *video* structures bring benefits they also introduce topical noise in the representation when used for television news programmes and keeping their influence constant between different document representations seems to be a beneficial strategy The benefits of the Jelmek-Mercer language model over the other smoothing methods is less severe for the semantic *shot+story* representation, which may indicate the potential for other smoothing methods to provide benefits in the semantic hierarchies

We have now completed our investigation of language models for the text-based retrieval of video shots and will proceed to investigate language models for the visual-based retrieval of video shots in the next chapter

CHAPTER VI

EVALUATION II. VIDEO RETRIEVAL USING VISUAL FEATURES

In this chapter we evaluate our discrete visual language modelling approach to video retrieval for the HSV colour, Canny edge, and DCT-based texture features on the TRECVid collections For each visual feature, we examine the effects of different smoothing techniques for both their global and regional histogram representations For HSV colour, we also investigate structural smoothing based on the video's physical and semantic composition structure, as examined for text based retrieval in the previous chapter Our approach in this chapter, is to examine each feature separately, while in the following chapter we will investigate the combination of these features' generative models into a single multimodal retrieval approach

6 1 Introduction

6 2 Overview of experiments

6 3 Experiments with colour features

6 4 Experiments with edge features

6 5 Experiments with texture features

6 6 Summary

6.1 Introduction

In the previous chapter we investigated video shot retrieval using text evidence alone in order to compare text-based language modelling approaches with non-language modelling text retrieval approaches for the video shot retrieval task and to quantify the performance of text-only approaches for video shot retrieval In a similar vein, in this chapter we compare language modelling approaches to standard visual retrieval ranking models, such as Manhattan distance, Euclidean distance, and Jensen-Shannon distance, for individual visual features and we establish the performance of the following separate visual features, HSV colour, Canny edges, and DCT-based texture, for the video shot retrieval-by-example search task

We experiment with regional versions of these features by including their relative X and Y position in the keyframe, scaled between 0 and 1, for each feature sample point We compare the retrieval performance for each feature using this regional representation quantised into 3x3, 4x4, and 5x5 regions per image in a multidimensional histogram representation

In this chapter we present results for each feature's performance independently of other visual and textual features We will deal with the problem of combining different features in the following chapter

The rest of this chapter is organised as follows in the next section we will present an overview of the experiments, followed by sections comparing our visual language modelling approach with

standard visual retrieval models for colour, edge and texture features

6.2 Overview of Experiments

We perform our experiments primarily on the TRECVid 2002 and TRECVid 2003 video shot search tasks but we treat each topic's image and video examples as a separate visual query-by-example topic We have only performed a select number of the visual experiments on the recent TRECVid 2004 collection The reported results, unless stated otherwise, are actually calculated on 70 visual query-by-example topics for TRECVid 2002 and 130 visual query-by-example topics for TRECVid 2003 instead of the standard 25 multimedia TRECVid topics for each collection The 70 and 130 visual examples consist of all the example images and videos that are part of the 25 TRECVid topic description for TRECVid 2002 and TRECVid 2003 respectively The TRECVid 2004 experiments when presented consist of either 23 multi-example topics or 140 single visual example topics

The visual experiments are divided into three sections the first deals with colour features, the second with edge features and the third with texture features For each feature we first present experiments on a *global* representation of the feature for the image in which we do not take any positional information into account We experiment with different levels of quantisation and in some cases different numbers of dimensions for the feature representation After establishing results for the global representation we experiment with representations that take into account positional information We perform experiments on these regional representations for 3x3, 4x4 and 5x5 regions on a single representation which we choose after first investigating the feature's global results

For each experiment we compare the language modelling approaches to the standard visual retrieval models, Manhattan distance, Euclidean distance, and Jensen-Shannon distance, which are described previously in chapter 3 We evaluate all language models we previously investigated in the text retrieval chapter except the Good-Turing, Dirichlet and Bayesian smoothing models Since each keyframe in the collection has a similar resolution, the Dirichlet smoothing reduces to Jelinek-Mercer smoothing, and therefore does not need to be tested Unfortunately, the Bayesian smoothing is too slow to compute as it typically involves 60,000 unique calculations to score each keyframe against a query image The Good-Turing smoothing performed poorly for text retrieval so we have not considered it for visual retrieval

As before, for the parametric language models we report the best results achieved by optimising the parameters and unbiased results in which the parameters have been optimised on one collection and tested on the other Since the two TRECVid collections are very different the unbiased results may in some cases be unrepresentative of a retrieval models performance This is due to differences in the collections which are at a high level, TRECVid 2003 is television news content from the late 1990's whereas TRECVid 2002 contains more general programmes predominately from between the 1950's and 1970's, and at a low-level, since the visual quality and consistency of the content is far superior in the TRECVid 2003 collection The supplemental results for TRECVid 2004 are tuned on TRECVid 2003

The video examples for TRECVid 2003 and 2004 are each supplied with a single keyframe,

which we use for the query image representation For TRECVid 2002 there is no supplied query-keyframes for the example videos, so we simply represent the query video using the nearest I-Frame to the middle of the query video We chose the middle I-Frame as opposed to the middle video frame as the I-Frame should be encoded with a higher visual quality It is interesting to review the topic images in relation to the relevant shots, as often it is the case that the query images and relevant shots are visually very dissimilar In chapter 3 Figures 17 and 18 (page 89) show the first topic's example images and sample relevant shots for the TRECVid 2002 and TRECVid 2003 search tasks For the experiments reported in this chapter the first topic in TRECVid 2002 represents 4 visual query-by-example topics and similarly the first topic in TRECVid 2003 represents 8 visual query-by-example topics

The experiments in this chapter use the supplied common keyframes for representing each shot in the TRECVid collections To generate this common shot boundary, small shots were combined with adjacent shots and in some cases inappropriate keyframes represent some of these sub-shot units In general, even without this combination of small shots, some shots may be better represented with multiple keyframes, however we have not yet compared the single keyframe approach with using multiple keyframes It is also important to note that for nearly all topics the relevant shot list contains shots which, by review of the supplied keyframe alone, would not be considered relevant by a human assessor For example at least two of the fifteen relevant shots for topic 75 (see Figure 17 on page 88) have keyframes that do not contain an image of the subject, Eddie Rickenbacker, of the search topic For topic 100 (see Figure 18 on page 89) at least 3 of the first 16 relevant shots have keyframes that do not contain the search topic 'aerial views of buildings or roads' These observations are typical of the shot keyframes and in fact other topics are far worse, such as all relevant shots for topic 119 for shots of Morgan Freeman have supplied keyframes in which Morgan Freeman is not visible

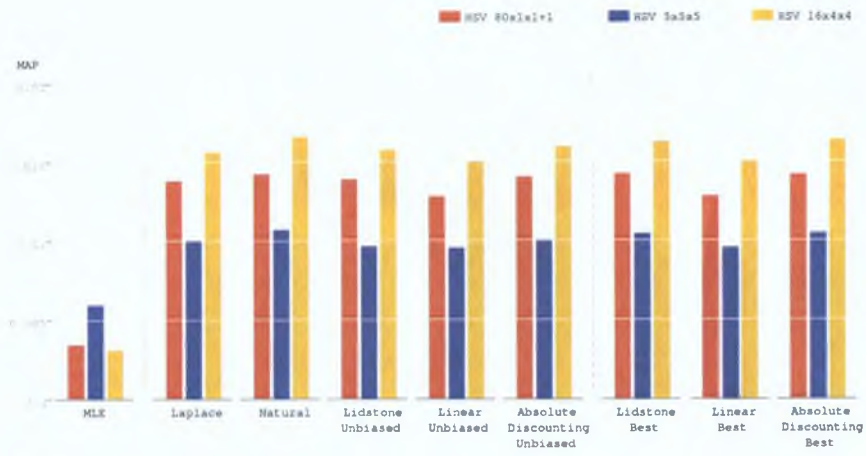
With these caveats noted about the limits of using single keyframe representations of shots for video retrieval, we will now present our query-by-example retrieval results for colour, edge and texture

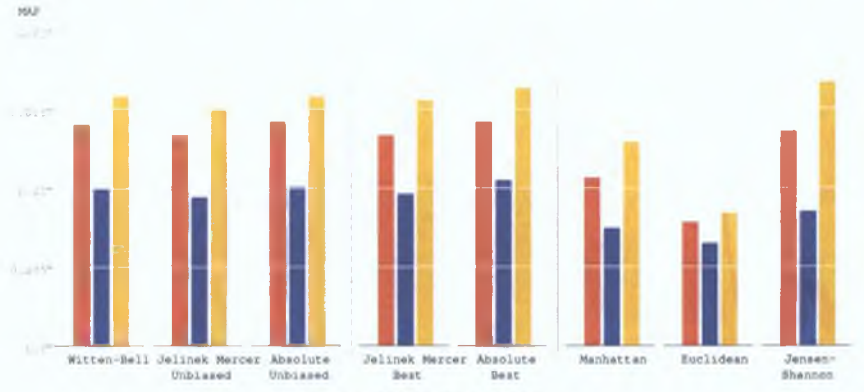
6.3 Experiments with Colour Features

In this section we investigate language models and standard visual retrieval models using the HSV colour feature for the TRECVid query-by-example task In the following subsections, we consider alternative quantisation of the 5 dimensional colour samples (X, Y, H, S, V), firstly into global colour representations and secondly into regional colour representations Finally, we consider smoothing using video structure both the physical and semantic structure that we investigated in the previous chapter in our text experiments

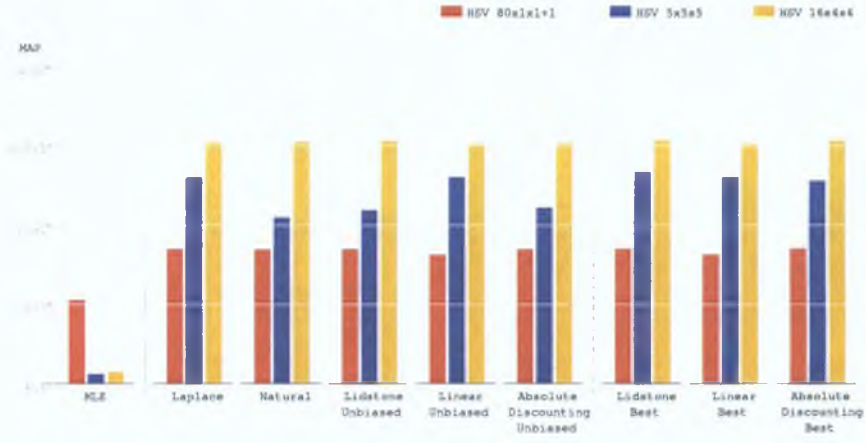
6 3 1 Global Colour

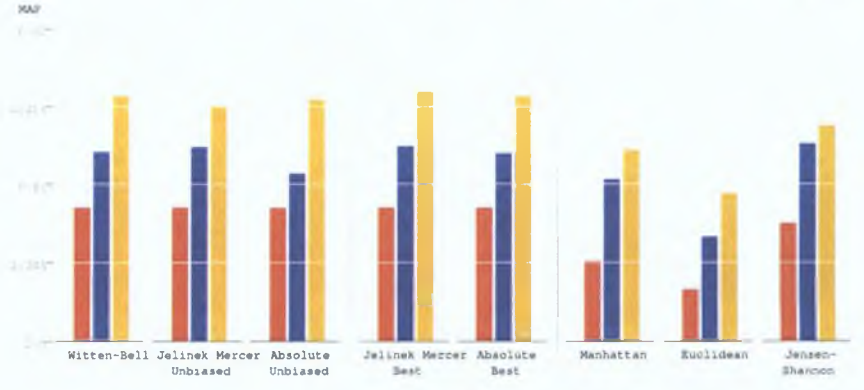
In this section we compare language models and standard visual retrieval models for the different *global* HSV colour representations *H 80+1*, *HSV 5x5x5* and *HSV 16x4x4*, which we described in Chapter 4 We establish a good representation of HSV colour information for the TRECVid query-by-example retrieval task that we will use in our regional colour experiments





(a) TRECVid 2002





(b) TRECVid 2003

Figure 30: Comparison of *global HSV histogram representations* (H 80 +1 desaturated, HSV 5x5x5, HSV 16x4x4) using language models and standard visual retrieval models for the *TRECVid 2002* and *TRECVid 2003* search task.

We display the results for these HSV representations for language models and visual retrieval models for the TRECVid 2002 and TRECVid 2003 search task in Figure 30 The *HSV 16x4x4* representation achieves on average 15% better MAP than the next best representation on both collections Its performance is followed by *H 80+1* and then *HSV 5x5x5* for the majority of retrieval models on the TRECVid 2002 search task For TRECVid 2003 the relative ordering of the different HSV representations is slightly different with the *HSV 5x5x5* representation second best and *H 80+1* the worst We believe that this change is caused by the difference in visual quality between the two collections The TRECVid 2003 collection possesses stable levels of saturation and brightness across encoded videos whereas within the TRECVid 2002 collection there are many differences in terms of brightness and saturation between the encoded videos, providing significant amount of unwanted noise in these bands relative to the hue dimension

For both TRECVid collections Jensen-Shannon distance is the best of the standard visual models across all HSV representations It achieves for the *HSV 16x4x4* representation a MAP of 0 0167 on TRECVid 2002 and a MAP of 0 0138 on TRECVid 2003 (see Table 43, appendix page 248, for *HSV 16x4x4* results) Manhattan is the second best standard visual retrieval model It is consistently better than Euclidean distance on both collections and on all the tested representations Its MAP is 0 0129, 22% lower than Jensen-Shannon distance, for TRECVid 2002 and 0 0122, 11% lower than Jensen-Shannon distance, for TRECVid 2003 for the HSV 16x4x4 representation

The MLE language model is, as expected, the worst performing retrieval model However, by smoothing the MLE model we achieve results nearly as good as the best standard visual model for TRECVid 2002 and better than the standard visual approaches for TRECVid 2003 Unlike text, for the HSV feature there is no significant difference between discounting and interpolation smoothing models for both collections

All the discounting methods have a similar performance except for Linear smoothing, which for most representations except *HSV 5x5x5* on TRECVid 2003, achieves the worst discounting model's MAP

The results for interpolation based language models are also indistinguishable For TRECVid 2002 Jelinek-Mercer smoothing performs slightly worse than Witten-Bell and Absolute interpolation smoothing which both have a MAP for *HSV 16x4x4* of 0 0158 compared to 0 0153, which could hardly be considered significant The results for TRECVid 2003 interpolation models have differences that are similarly insignificant

The optimum parameters of the parametric language models for TRECVid 2002 and TRECVid 2003 while different (specifically look at results for *HSV 16x4x4* in Table 43, appendix page 248) achieve similar results for the unbiased runs This may indicate that the performance of language models for the HSV colour feature is somewhat resilient to slightly different amounts of smoothing for the information retrieval task It is also true that global features are dense histograms and therefore require only a small amount of smoothing

6 3 2 Regional Colour

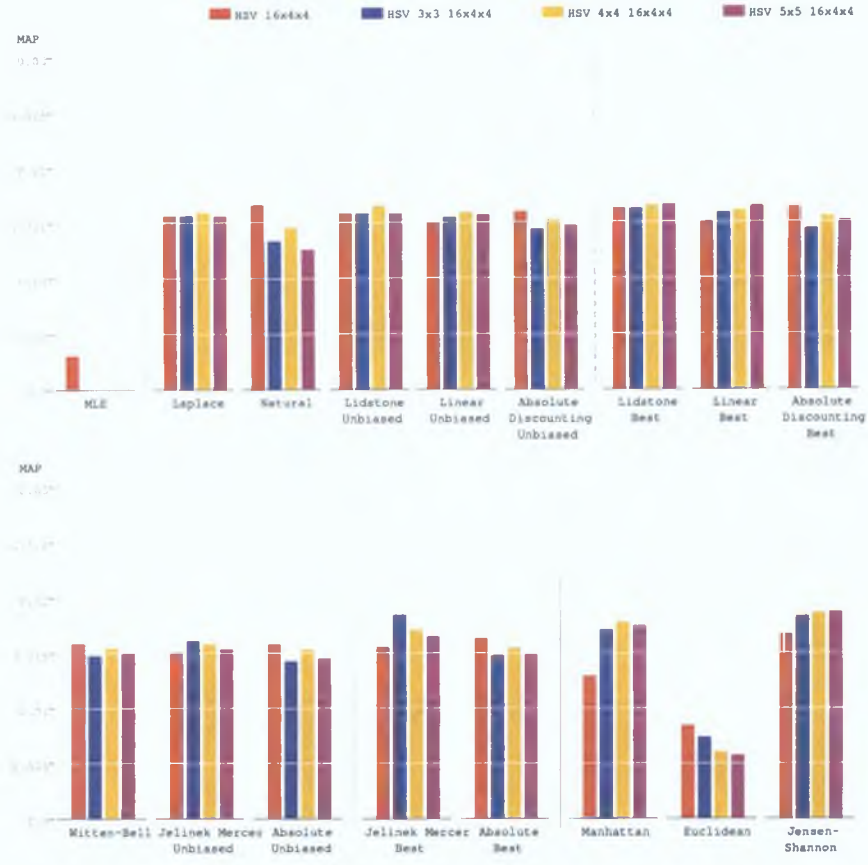
In this section we investigate language models and standard visual retrieval models using the *regional* HSV colour feature for the TRECVid query-by-example task We use regional variations of the *HSV 16x4x4* representation as this performed consistently well in the *global* colour experiments In our experiments we quantise the X, Y position dimensions for each HSV sample uniformly into 3, 4 or 5 levels thereby essentially breaking the image into 9, 16, or 25 independent rectangular regions and producing a language with 2304, 4096, and 6400 symbols respectively

We present the results in Figure 31 for the 3x3, 4x4 and 5x5 regional *HSV 16x4x4* representations for the TRECVid 2002 and TRECVid 2003 search task The results for TRECVid 2002 are mixed showing small increases and decreases in MAP for different retrieval models for 3x3 regions compared to the global representation and showing only marginal improvements for 4x4 and 5x5 regions In contrast, the results for TRECVid 2003 show major improvements in all retrieval models (except MLE) for all regional representations compared to the global representation The retrieval results for 3x3, 4x4 and 5x5 regions increase the MAP by respectively 53%, 66% and 86% compared to the MAP of the global HSV representation

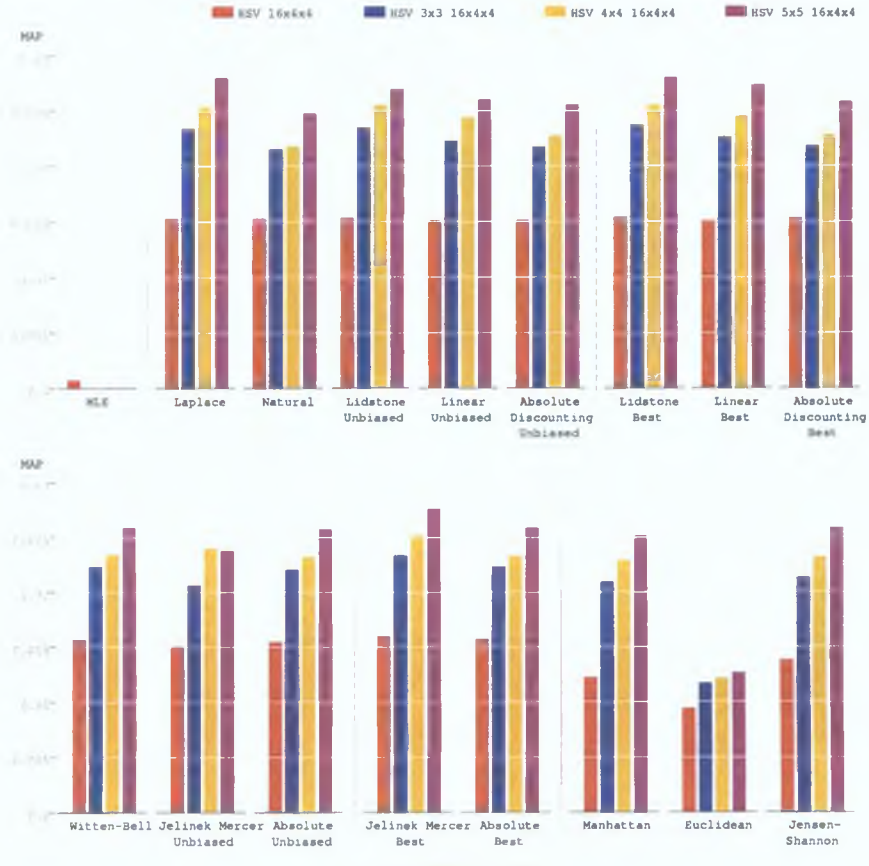
The relative ordering of the standard visual models remains unchanged for all tested regional colour representation on both collections Jensen-Shannon distance is the best, Manhattan distance is the next best and Euclidean distance is the worst For each standard visual model, except Euclidean distance on the TRECVid 2002 collection, the use of regions improves upon their global results For TRECVid 2003 Euclidean distance improves when using regions but it still achieves only 50% of the other visual model's MAP Jensen-Shannon distance achieves a maximum MAP of 0 0187 for 5x5 regions, a 12% improvement on its global results on TRECVid 2002 For TRECVid 2003 it achieves a maximum MAP at 5x5 regions of 0 0258, which is a 87% improvement on global results (see Table 44, appendix page 249) For both collections the difference in MAP for Jensen-Shannon distance compared to Manhattan distance is reduced when using regional colour compared to global colour For example for TRECVid 2003 Jensen-Shannon distance is only 1% better than Manhattan model for 5x5 regional representation compared to an 11% difference when comparing them on the global colour representation

For language models, the MLE model achieves a MAP of zero for regional colour representations This is due to a far larger vocabulary of symbols in the regional languages, which increases the number of symbols with a zero frequency Smoothing the MLE, of course, solves this zero frequency problem, which is more severe in a visual query-by-example task than for a short text query

The discounting language models show only tiny improvement with the addition of regions for TRECVid 2002 In fact, Natural and Absolute discounting have losses in MAP Lidstone discounting remains roughly unchanged for global and regional representations In contrast, for TRECVid 2003 all discounting models show improvements in MAP The Laplace and Lidstone discounting models *are the best overall retrieval models* with a MAP of 0 0280 for 5x5 regions (81% improvement compared to global representation), which is achieved even for the unbiased Lidstone parameters (see Table 44, appendix page 249) It is surprising that the somewhat arbitrary just-add-one rule of Laplace smoothing is so effective in this case As well as having a Bayesian interpretation, Laplace and Lidstone can be interpreted as interpolation with a uniform



(a) TRECVid 2002



(b) TRECVid 2003

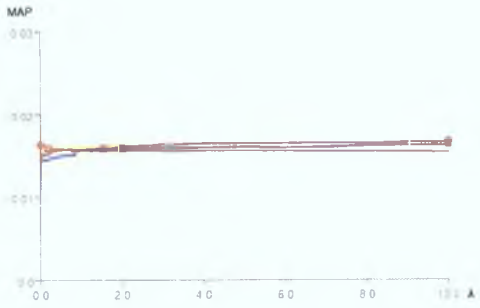
Figure 31: Comparison of *regional histogram representations* (HSV 16x4x4 for no regions, 3x3, 4x4, and 5x5 regions) using language models and standard visual retrieval models for the *TRECVid 2002* and *TRECVid 2002* search tasks.

source.

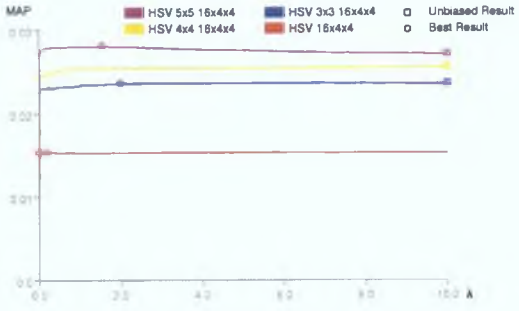
The related Jelinek-Mercer smoothing, interpolation with the background collection model, is the next overall best retrieval model on the TRECVid 2003 collection with a maximum MAP of 0.0276 but unfortunately for unbiased parameters achieves only a MAP of 0.0238, which is the second lowest result for 5x5 regions. On both collections the other interpolated smoothing models Witten-Bell and Absolute smoothing perform slightly worse than Jelinek-Mercer smoothing except for the unbiased result for Jelinek-Mercer for 3x3 and 5x5 regions on the TRECVid 2003 collection. For TRECVid 2002 the Witten-Bell and Absolution interpolation-based smoothing actually decrease in effectiveness with the addition of regions in this experiment. Jelinek-Mercer interpolation smoothing shows significant improvement from no regions (MAP 0.0155) to 3x3 regions (MAP 0.0185) with a 19% improvement. This improvement does not extend to more regions and the MAP steadily decreases for 4x4 and 5x5 regions. For regional HSV colour there does not seem to be much advantage, if any, for using a background collection model for smoothing - discounting methods achieve just as good, if not better, results. With the results also indistinguishable for the global colour representation, we believe that the HSV colour visual language is just as well smoothed with a uniform distribution as with the background collection distribution. This is in contrast to the more semantic and skewed text language models.

Overall the unbiased results for the parametric language models are very close to the maximised results for both collections (see Figure 32). The Jelinek-Mercer language model is the exception for both collections (see Figures 32(g) and 32(h)) – for regions in the TRECVid 2002 search task this method requires a lot of smoothing with the background collection ($\lambda = 0.65$ for 3x3 and 5x5 regions) but for TRECVid 2003 this methods requires a small magnitude of smoothing ($\lambda = 0.05$ for any regional representation). These parameter values are at different ends of the scales and result in poorer results when exchanged to produce the unbiased results. This instability could be due to our tuning procedure or is maybe inherent in the application of the Jelinek-Mercer language model to this feature.

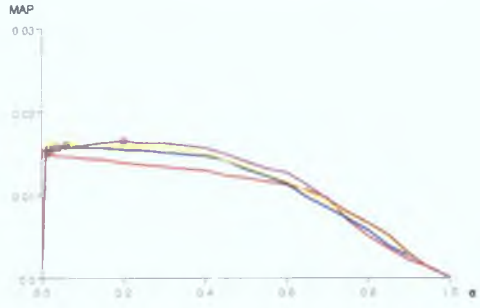
We present statistical significance comparisons of the unbiased retrieval models on the *HSV 5x5 16x4x4* feature for the set of topic results across the three TRECVid collections in Table 11. This table indicates that the Laplace language model achieves the best overall MAP result for the regional colour feature, though the Jensen-Shannon distance is only marginally lower (0.3% difference). Jensen-Shannon distance is actually the stronger result and is statistically significantly better than all other retrieval models below it. The unbiased Lidstone result achieves a similar MAP to the related Laplace smoothing with a statistically insignificant 2.1% difference. Jelinek-Mercer smoothing achieves the worst overall MAP (we've excluded the poorer MLE and Euclidean models from this comparison to save space) but is notably not statistically significantly poorer than any other smoothed visual language modelling, though it is statistically significantly lower than Jensen-Shannon and Manhattan distance. It must be remembered that the magnitude of the MAP differences are very small and are maybe not observable by real users. The Jensen-Shannon distance, which is marginally better than the language modelling approach, is a similar approach to the language modelling approach but combines both the query and document into the hypothetical source before measuring the relative entropy from the query and document to it. The two best performing retrieval models are non-parametric and therefore it is possible that improvements to our parameter tuning procedure could redress some of the slight performance differences with the parametric language models.



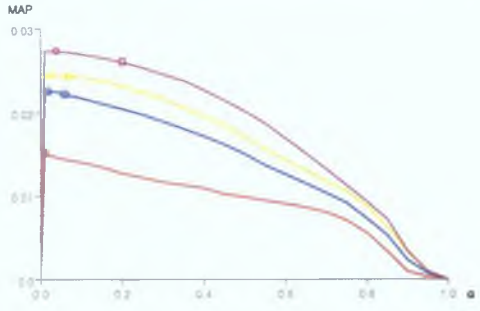
(a) TRECVid 2002: Lidstone



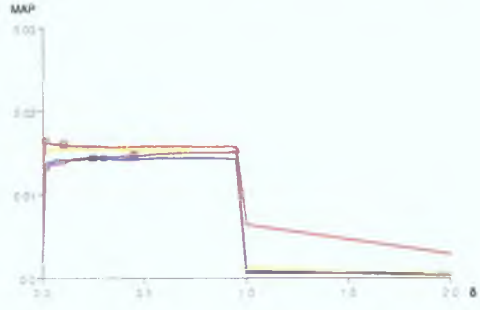
(b) TRECVid 2003: Lidstone



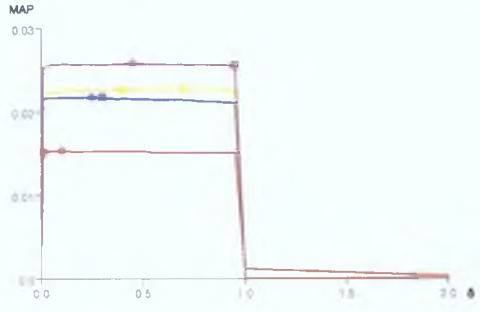
(c) TRECVid 2002: Linear



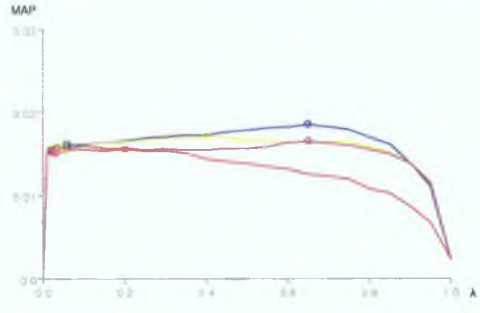
(d) TRECVid 2003: Linear



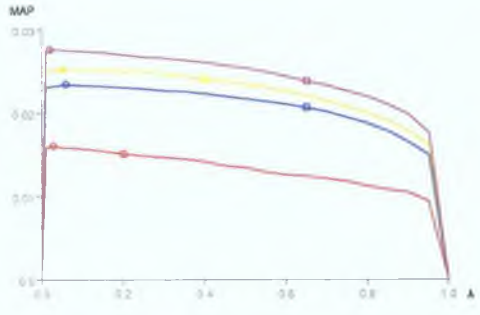
(e) TRECVid 2002: Absolute Discounting



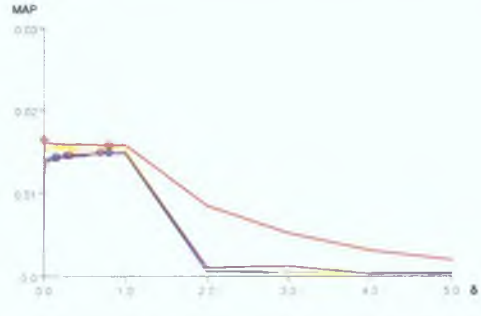
(f) TRECVid 2003: Absolute Discounting



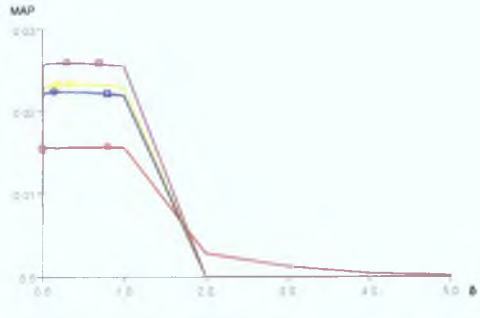
(g) TRECVid 2002: Jelinek-Mercer



(h) TRECVid 2003: Jelinek-Mercer



(i) TRECVid 2002: Absolute Interpolation



(j) TRECVid 2003: Absolute Interpolation

Figure 32: Plot of MAP over the parameter space for the parametric language models using regional colour representations for the *TRECVid 2002* search task and *TRECVid 2003* search task.

Table 11: Statistical significance comparison of retrieval models on the regional 5x5 HSV 16x4x4 colour feature for the aggregated TRECVid 2002, 2003 and 2004 search tasks.

Ret. Meth.	*MAP/%Dif*	*Wilcoxon Test Results*								
Laplace (Lap)		-	>JS	>Lid	>Lin	>WB	>Man	>Abs	>AbsD	>JM
(Lap)	0.0176		.997	.950	.040	.000	.982	.000	.000	.466
Jensen-Shannon (JS)		<Lap	-	>Lid	>Lin	>WB	>Man	>Abs	>AbsD	>JM
(JS)	0.0176/-0.3%	.997		.000	.000	.000	.007	.000	.000	.000
Lidstone (Lid)		<Lap	<JS	-	>Lin	>WB	>Man	>Abs	>AbsD	>JM
(Lid)	0.0172/-2.1%	.950	.000		.000	.000	.994	.000	.000	.166
Linear (Lin)		<Lap	<JS	<Lid	-	>WB	>Man	>Abs	>AbsD	>JM
(Lin)	0.0169/-4.2%	.040	.000	.000		.000	.988	.000	.000	.649
Witten-Bell (WB)		<Lap	<JS	<Lid	<Lin	-	>Man	>Abs	>AbsD	>JM
(WB)	0.0166/-5.9%	.000	.000	.000	.000		.997	.001	.091	.999
Manhattan (Man)		<Lap	<JS	<Lid	<Lin	<WB	-	>Abs	>AbsD	>JM
(Man)	0.0165/-6.1%	.982	.007	.994	.988	.997		.002	.001	.012
Absolute (Abs)		<Lap	<JS	<Lid	<Lin	<WB	<Man	-	>AbsD	>JM
(Abs)	0.0164/-6.6%	.000	.000	.000	.000	.001	.002		.389	1.000
Absolute Disc. (AbsD)		<Lap	<JS	<Lid	<Lin	<WB	<Man	<Abs	-	>JM
(AbsD)	0.0164/-6.7%	.000	.000	.000	.000	.091	.001	.389		1.000
Jelinek-Mercer (JM)		<Lap	<JS	<Lid	<Lin	<WB	<Man	<Abs	<AbsD	-
(JM)	0.0159/-9.7%	.466	.000	.166	.649	.999	.012	1.000	1.000	

We tabulate the results of our tested retrieval models on the official TRECVid multi-example topics in Table 12. Previous to this our results have been for single example searching and therefore could not be directly compared with official TRECVid results. The single topic results were fused using CombSUM with varying amounts of truncated results (tuned using the normal procedure) to produce these multi-example results. The results are very close for all retrieval models except Euclidean and MLE (not shown). The results are better than the continuous GMM approach using the DCT feature (see Table 2, page 79) for the TRECVid 2003 and TRECVid 2004 collection but worse for TRECVid 2002 search task. The difference in visual quality between TRECVid 2002 and the other collections may account for difference in results. Query keyframes for the example videos were not distributed with the TRECVid 2002 collection, so it is possible their selection may be accountable for the negative difference on this collection. The relative difference between the Lidstone language model and the best GMM query-likelihood approach are respectively -49.5%, 42.3%, and 75.0% for the three search tasks. Our other discrete visual language models have similar relative performance. The popular standard retrieval model Euclidean distance performs very poorly relative to the GMM and discrete visual language modelling approaches. Euclidean distance magnifies the difference between histogram bins, which is a fine strategy for exact match searching but not for the more general TRECVid search task.

6.3.3 HSV colour experiments using the physical and semantic video structure

In this section we investigate the usefulness of the physical and semantic video composition structure of edited video content in our discrete visual language modelling approach for the colour feature. Smoothing using the video structure of video is useful for text search. In this

Table 12: Comparison of retrieval models on the HSV 5x5 16x4x4 feature for official TRECVid topics (i.e. fused topic examples).

VisExsCol-CombScore	*TRECVid 2002*				*TRECVid 2003*				*TRECVid 2004*				*TRECVid 02-04*			
Retrieval Method	*MAP*	*P10*	*P30*	*P100*	*MAP*	*P10*	*P30*	*P100*	*MAP*	*P10*	*P30*	*P100*	*MAP*	*P10*	*P30*	*P100*
Laplace	**.0150**	.052	.035	.022	**.0394**	.080	.091	.058	**.0169**	.065	.049	.036	**.0240**	.066	.058	.039
Best	**(.0150**	.056	.036	.023)	**.0395**	.080	.091	.058	**(.0179**	.052	.051	.037)	**(.0243**	.063	.059	.039)
Lidstone	**.0145**	.048	.036	.024	**.0411**	.080	.097	.060	**.0175**	.057	.051	.036	**.0246**	.062	.062	.040
Best	**(.0153**	.048	.036	.023)	**(.0401**	.080	.092	.058)	**(.0182**	.052	.051	.037)	**(.0247**	.060	.060	.039)
Linear	**.0148**	.052	.036	.021	**.0392**	.080	.092	.058	**.0172**	.070	.049	.037	**.0239**	.067	.059	.039
Best	**(.0140**	.048	.035	.022)	**(.0381**	.088	.085	.056)	**(.0173**	.057	.052	.036)	**(.0233**	.064	.058	.038)
Absolute Discounting	**.0141**	.052	.035	.019	**.0340**	.080	.073	.050	**.0149**	.057	.046	.033	**.0211**	.063	.052	.034
Best	**(.0150**	.052	.033	.020)	**(.0336**	.084	.073	.051)	**(.0150**	.061	.042	.033)	**(.0214**	.066	.050	.035)
Witten-Bell	**.0149**	.052	.037	.022	**.0353**	.080	.076	.054	**.0164**	.057	.049	.034	**.0223**	.063	.054	.037
Best	**(.0149**	.052	.039	.022)	**(.0354**	.080	.076	.055)	**(.0166**	.057	.051	.034)	**(.0224**	.063	.055	.037)
Jelinek-Mercer	**.0152**	.044	.036	.023	**.0398**	.068	.081	.060	**.0174**	.052	.054	.036	**.0243**	.055	.057	.040
Best	**(.0142**	.036	.027	.020)	**(.0390**	.080	.083	.058)	**(.0175**	.057	.054	.036)	**(.0237**	.058	.054	.038)
Absolute	**.0144**	.048	.037	.022	**.0349**	.076	.075	.054	**.0152**	.052	.043	.032	**.0217**	.059	.052	.036
Best	**(.0156**	.052	.040	.022)	**(.0345**	.092	.075	.052)	**(.0154**	.061	.045	.032)	**(.0220**	.068	.053	.036)
Manhattan	**.0183**	.040	.029	.025	**.0366**	.108	.093	.060	**.0179**	.061	.052	.032	**.0245**	.070	.058	.039
Best	**(.0186**	.048	.027	.026)	**(.0373**	.108	.095	.060)	**(.0195**	.065	.054	.033)	**(.0253**	.074	.058	.040)
Jensen Shannon	**.0190**	.036	.040	.025	**.0365**	.100	.091	.058	**.0223**	.091	.057	.036	**.0260**	.075	.063	.040
Best	**(.0192**	.036	.039	.024)	**(.0366**	.100	.091	.058)	**(.0244**	.078	.057	.039)	**(.0268**	.071	.062	.040)
Euclidean	**.0064**	.024	.028	.018	**.0161**	.044	.049	.036	**.0069**	.022	.028	.020	**.0099**	.030	.035	.025
Best	**(.0064**	.024	.028	.018)	**(.0161**	.044	.049	.036)	**(.0069**	.022	.028	.020)	**(.0099**	.030	.035	.025)
Average of unbiased	**.0146**	.045	.035	.022	**.0353**	.080	.082	.055	**.0163**	.058	.048	.033	**.0222**	.061	.055	.037
Average of best	**.0148**	.045	.034	.022	**.0350**	.084	.081	.054	**.0169**	.056	.048	.034	**.0224**	.062	.055	.037

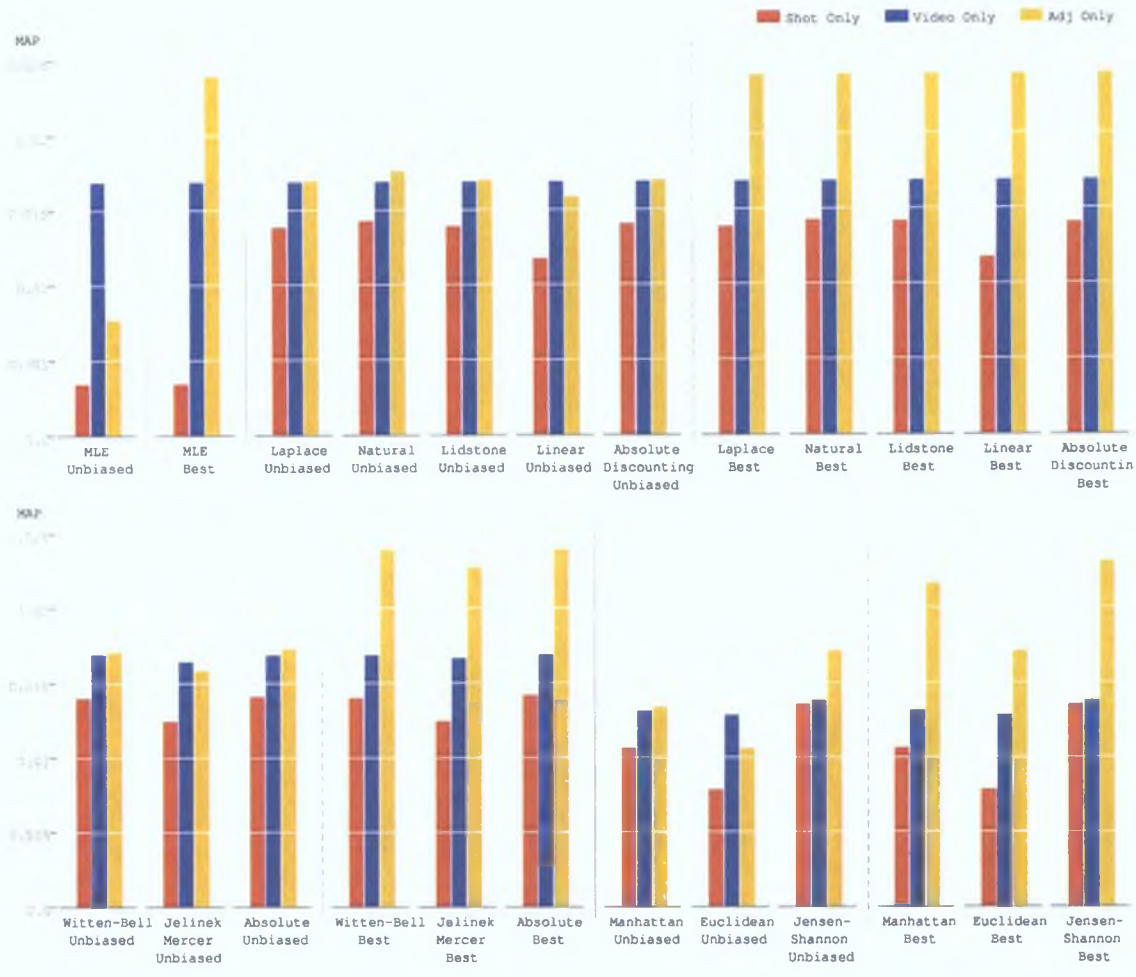


Figure 33: Comparison of the *indexing units* shots, sequence of adjacent shots, and videos with the HSV 80x1x1+1 colour representation for language models and visual retrieval models on the *TRECVid 2002* search task.

section we will evaluate the following structural units:

- adjacent shots: the colour model for a window of shots that spans a fixed number of shots before and after the given shot,
- video: the colour model for the physical video that contains the given shot (i.e. colour model of the entire video programme),
- story: the colour model for the semantic story that contains the given shot.

We will first investigate the adjacent shots, video and story structures separately as our representation or indexing unit for each shot for the TRECVid search tasks and following this we will experiment with structural smoothing, combining the structural models with the shot model.

The results for language models and standard visual retrieval models using different structural units are shown in Figures 33 and 34 for the TRECVid 2002 and TRECVid 2003 search tasks.

Without exception for TRECVid 2002, see Figure 33, every language model and standard retrieval model achieves a higher MAP when using the adjacent shots or the video representation

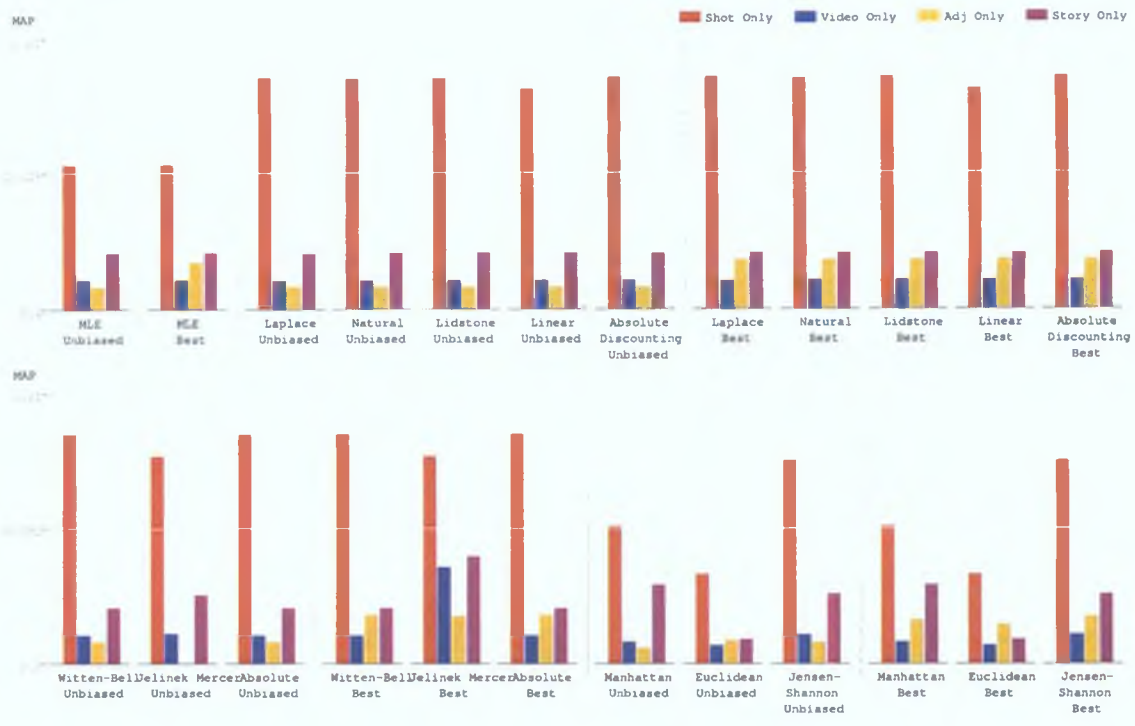


Figure 34: Comparison of the *indexing units* shots, sequence of adjacent shots, videos, and stories with the HSV 80x1x1+1 colour representation for language models and visual retrieval models on the *TRECVid 2003* search task.

than when using the shot representation alone. For TRECVid 2003, see Figure 34, the video, adjacent shots and story structures are all clearly inferior to the shot representation alone for all the evaluated retrieval models. We believe that this contradiction between the results for the two collections is because most of the videos in TRECVid 2002 are about a single topic, unlike the news videos in TRECVid 2003, and as a consequence the relevant shots for many of the topics in TRECVid 2002 are more clustered together in the videos whereas for TRECVid 2003 the relevant shots are more spread across multiple news broadcasts. Some of the videos in the TRECVid 2002 collection have very distinctive encodings and some query examples were chosen from the test collection, and this possibly accounts for the benefit of the video structure for this collection.

The results for the TRECVid 2002 search task, tabulated in Table 45 (appendix page 250), show that while the improvement in MAP for the adjacent shots and video structure is large, it is not statistically significant according to the Wilcoxon rank test. The high Wilcoxon p-values, greater than 0.5 for video representation and greater than 0.25 for adjacent shots representation, indicate an inconsistent improvement in MAP. This increase in performance from using the video and adjacent shots structure are possibly topic dependent and may also require that the topic image has an effective shot-only performance before it can be effective with the shot or video structure, in much the same way as the use of relevance feedback would require a reasonable initial retrieval performance. Since most topic images have very poor performance to begin with, the performance for these unsuccessful images will likely further deteriorate when using the video and adjacent shots structure.

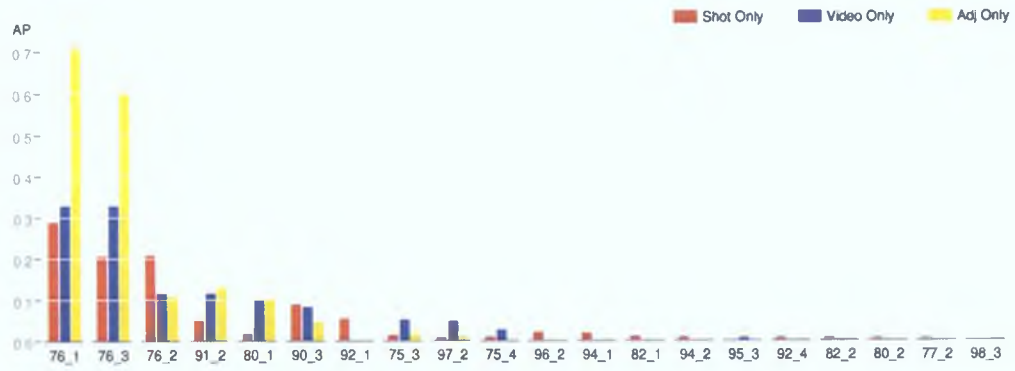


Figure 35: Comparison of the *indexing units* shots, sequence of adjacent shots, videos and stories with the HSV 80x1x1+1 colour representation on the 20 most successful *TRECVid 2003* topic images using the best Lidstone language models

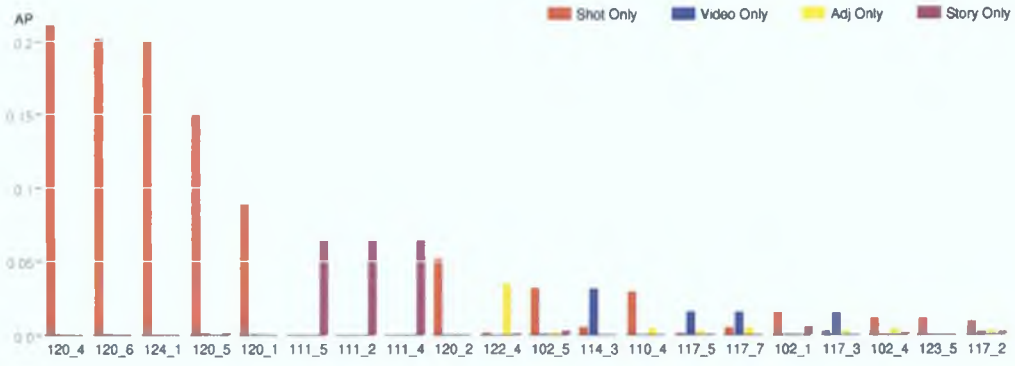


Figure 36: Comparison of the *indexing units* shots, sequence of adjacent shots, videos and stories with the HSV 80x1x1+1 colour representation on the 20 most successful *TRECVid 2003* topic images using the best Lidstone language models

We will investigate this difference between both collections more closely by looking at the results for the 20 best topic images on both collections. These results using the optimised Lidstone smoothed language model for the TRECVid 2002 and TRECVid 2003 search tasks are shown in Figures 35 and 36.

The results for the top 20 topic images for TRECVid 2002 in Figure 35 indicate that the performance boost in the MAP measure for the adjacent shot structure is mainly attributable to two topic images, $topic76_1$ and $topic76_3$, both from the TRECVid topic 76. In fact, only 7 of the top 20 topic images show improvement in the adjacent shots representation over the shot-only representation and 13 show a decrease in average precision, which explains how the MAP measure can have a large increase but simultaneously be statistically insignificant. The average precision for the video representation increases for 8 of the top 20 topic images, while 12 of the top images show decreases in average precision. Similar to the adjacent representation, the increases for the video representation are mainly from the best performing topic images within the top 20 topic results. We see that for TRECVid 2002 the majority of topic images are better represented or indexed using the shot representation than with the adjacent shots or video representation even though a naive look at the MAP score alone might lead to the opposite conclusion.

The results for the top 20 topic images for TRECVid 2003 in Figure 36 show that the top 5 topic images, mostly from topic 120, achieve their best results using the shot-only representation In fact the shot-only representation is best for 12 of the 20 best topic images The story representation produces the next 3 best topics images, which are all from the TRECVid topic 111 The adjacent shots representation is best for only one of the topic images and the video representation is best for only 4 of the top 20 topic images The consistent message from both collections is that the shot representation is better than the other tested representations for colour except for a small minority of topic images

As can be seen in Table 45 (appendix page 250) for TRECVid 2002 the video representation improves on the shot representation by about 20% for the retrieval models in terms of MAP The MLE language model for the video structure achieves the best MAP of 0 0169 The other smoothed language models *do not improve* on the MLE model The MLE model also outperforms the results of Jensen-Shannon distance, which produced the best MAP of 0 0138 for the standard visual models

For TRECVid 2003, see Figure 34, the physical video structure performs 8 times worse than the shot representation for most of the retrieval models The semantic story structure performs twice as well as the video structure producing MAP scores for the retrieval models that are a quarter of the shot-only representation scores

For the adjacent shots representation, the optimum number of adjacent shots before and after the current shot is very different for both collections For TRECVid 2002, 20 shots before and after the indexed shot are best - more may actually be better as 20 was the upper limit for optimising this parameter For TRECVid 2003 one extra shot before and after the indexed shot is best for the adjacent shots structure The results for using adjacent shots for TRECVid 2003, see Table 46, significantly underperform the shot representation with the MAP measure decreasing by 80% for the best discounting models on this representation In fact we could say 0 adjacent shots (the shot model itself) is the best adjacent shot model for TRECVid 2003 In contrast, for TRECVid 2002, see Table 45, even the unbiased adjacent shots structure for a window of +/- 1 adjacent shots helps, increasing the MAP by 20%, but this is considerably smaller than the potential in TRECVid 2002 when the adjacent shots structure uses +/- 20 shots for its window width, which increases MAP by 70% to 0 0239 for all language models For the TRECVid 2002 search task an adjacent window of any size from 1 to the maximum number of shots in a video will produce better results than a shot-only representation The opposite is true for TRECVid 2003 However, as mentioned previously, these results concerning the adjacent shots structure are due to a small number of topic images and are not representative of the true effects these structures have on the majority of topic images

We now consider using these structures for hierarchical smoothing of the visual shot language model In the previous chapter we saw how the hierarchical smoothing structures can achieve consistent improvements over the shot text model alone We present the results for *structural smoothing* for the TRECVid 2002 and TRECVid 2003 search tasks in Figures 37 and 38

The results for TRECVid 2002 in Figure 37 show that Jelinek-Mercer smoothing is the only structural smoothing language model that improves the performance over the shot-only representation when using the structures *shot+video* and *shot+adj+video* The other interpolation-based

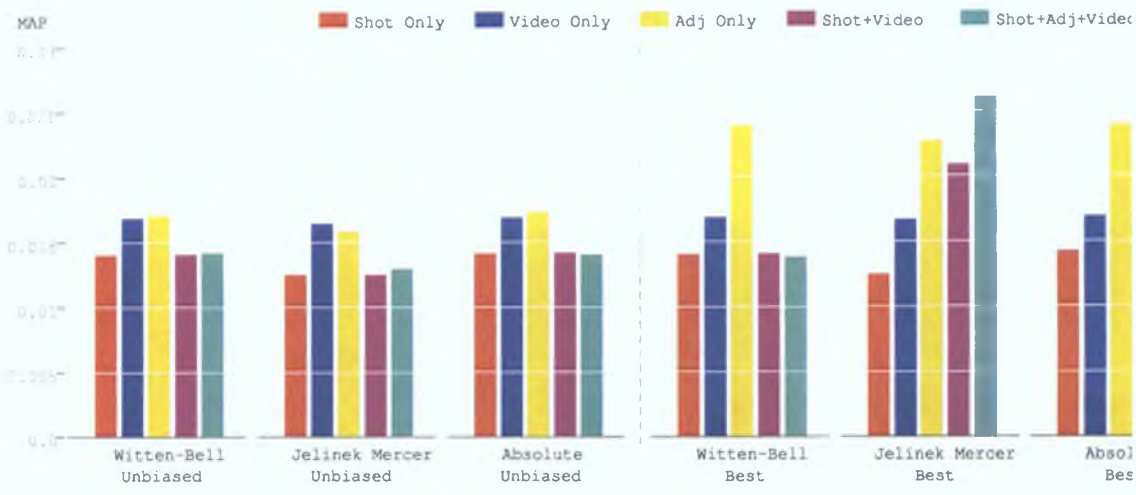


Figure 37: Comparison of *smoothing with structural units* (shots+videos, shots+sequence of adjacent shots+videos) and *smoothing different indexing units* (shot-only, video-only, adjacent shots-only) with the HSV 80x1x1+1 colour representation for interpolation-based language modelling information retrieval on the *TRECVid 2002* search task.

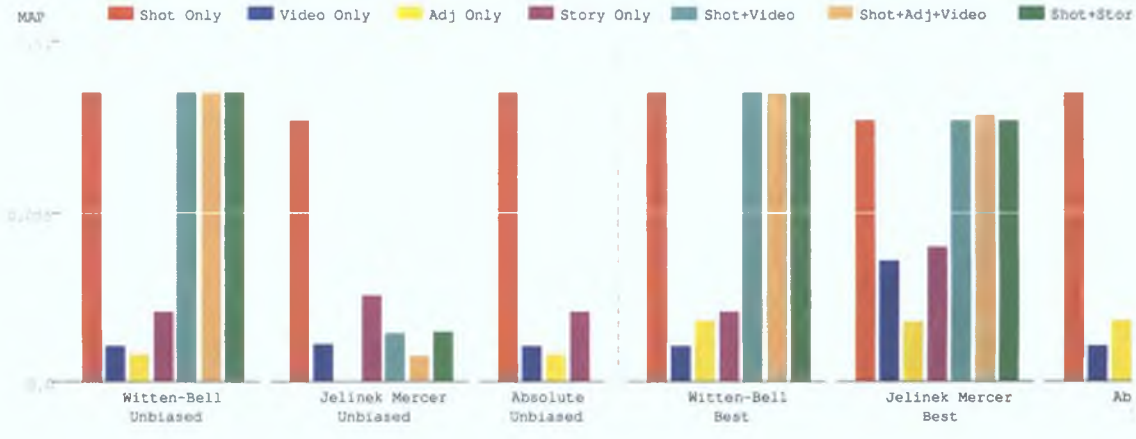


Figure 38: Comparison of *smoothing with structural units* (shots+videos, shots+adjacent shots+videos) and *smoothing different indexing units* (shot-only, video-only, adjacent shots-only) with the HSV 80x1x1+1 colour representation for interpolation-based language modelling information retrieval on the *TRECVid 2003* search task.

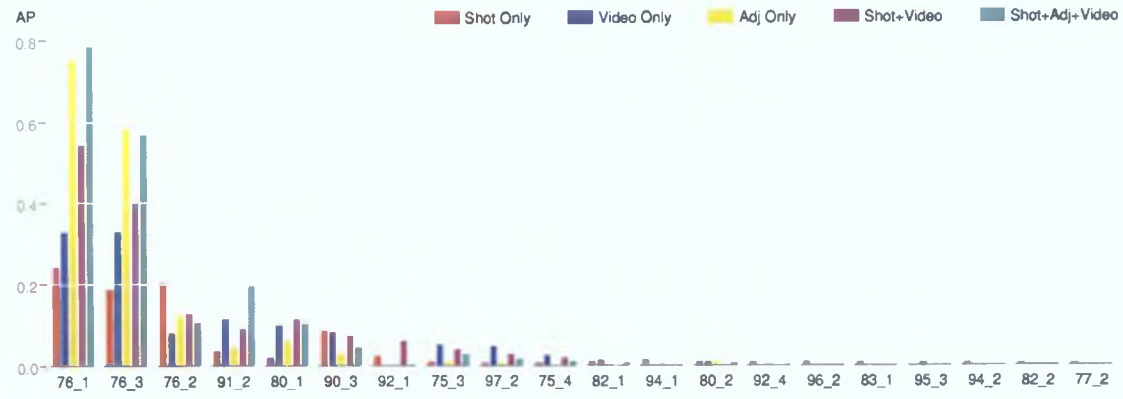


Figure 39: Comparison of *smoothing with structural units* (shots+videos, shots+adjacent shots+videos) and *smoothing different indexing units* (shot-only, video-only, adjacent shots-only) with the HSV 80x1x1+1 colour representation on the 20 most successful *TRECVid 2002* topic images using the best Jelinek-Mercer structural smoothing model.

smoothing models do not improve on the results compared to the smoothed shot-only models.

The results for TRECVid 2003 in Figure 38 are unsurprisingly very negative for smoothing the keyframe colour model with the physical structure combinations *shot+video* and *shot+adj+video* or with the semantic *shot+story* composition structure. These structure-based estimators provide no benefit in the TRECVid 2003 collection - the maximum possible MAP does not even improve compared to the smoothed shot-only models. Due to this failure the unbiased results for TRECVid 2002, which use the best parameters from TRECVid 2003, lack significance. The shots within television news stories are visually dissimilar moving from anchorperson shots to other types of video sequences such as interviews or external reports and therefore a simple model of colour across the whole story is not much use. For the rest of this section we will look further at the TRECVid 2002 results.

Figure 39 shows the average precision results for the 20 best topic images for structural smoothing compared with smoothing the shot-only, adjacent shots-only and video-only models on the TRECVid 2002 collection using Jelinek-Mercer smoothing. The *shot+adj+video* structure improves upon the adjacent shot-only representation for 8 out of the 20 best topic images but it has the best average precision for only 2 of the topics, $topic76_1$ and $topic91_2$. The *shot+video* structure improves on both the shot-only and video-only representation for 4 topic images but for topic images below the top 10 it achieves very poor results.

The results for structural-based smoothing on the TRECVid 2002 collection are shown in Table 13 for the *shot+video* hierarchical structure and in Table 14 for the *shot+adj+video* structure. The interpolated Jelinek-Mercer smoothing has the best potential results with 109.7% improvement for the MAP measure for the *shot+adj+video* structure and has a 67.9% improvement for *shot+video* structure compared to smoothing the shot with only the collection model as in the shot-only language models. For the unbiased results this potential performance gain is not achieved. The results for optimised structural smoothing are better than using the structural units separately as presented previously in Table 45. The improvements for both *shot+video* and *shot+adj+video* structures over the shot-only representation are not statistically significant with Wilcoxon p-values of 0.628 and 0.839 respectively. Unlike Jelinek-Mercer smoothing of hierarchical structure, if we use the hierarchical version of Witten-Bell and Absolute interpolation

Table 13: Results for smoothing the HSV 80x1x1+1 *shot+video* hierarchical colour structure for the *TRECVid 2002* search task.

TRECVID 2002	*Shot+Video*					*V. Shot Only*	*V. Video Only*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*
Witten-Bell		**.0140**	.034	.023	.017	+0.4% ∼ .203	-16.7% ∼ .598
Jelinek Mercer	$\lambda_{shot}, \lambda_{vid}, \lambda_{col}$						
Unbiased	0.95, 0.00, 0.05	**.0124**	.025	.024	.018	-0.0% ∼ .495	-24.3% ∼ .604
Best	(0.05, 0.95, 0.00	**.0209**	.016	.023	.023)	+67.9% ∼ .628	+25.2% ∼ .741
Absolute	$\delta_{shot}, \delta_{vid}$						
Unbiased	0.15, 0.00	**.0141**	.034	.024	.017	+0.4% ∼ .070	-16.3% ∼ .601
Best	(0.02, 0.00	**.0142**	.036	.024	.017)	-0.3% ∼ .934	-16.1% ∼ .596
Average of unbiased		**.0135**	.031	.024	.017	+0.3%	-19.1%
Average of best		**.0175**	.026	.024	.020	+33.8%	+4.5%

to smooth the same structural units we do not improve on the results in terms of the MAP measure compared to simply smoothing the shot model with the collection model.

In summary, interpolation-based smoothing using the video's hierarchical structure is a successful strategy for the TRECVid 2002 search task for a minority of high performing topic images but fails for the TRECVid 2003 search task. We found that for news content in TRECVid 2003 and for the majority of topic images for TRECVid 2002's general video collection that structural units at a higher level than shot such as story, fixed sized adjacent shot windows and videos provide poor performance for visual features. It may be that sub-story units would be useful for visual structural smoothing. However, segmentation of stories, for example into anchorperson, interview, and report segments, was unavailable to us. Another strategy would be to use structural smoothing within a shot to combine the keyframe visual feature model with a shot model (generated from all its frames or I-Frames) and possibly with an adjacent frame model. Due to the very poor performance of structural smoothing involving videos, stories and adjacent shots for the colour feature on TRECVid 2003, we will not pursue it further in this chapter for the other visual features.

6.4 Experiments with Edge Features

In this section we investigate visual language models for the Canny edge feature (Canny, 1986) for the TRECVid visual search task. We first compare the language models on the global edge representation with from 4 to 64 levels of quantisation and with an extra symbol for the out-of-bounds edge direction, producing a language consisting of between 5 and 65 symbols. After establishing a reasonable global representation of edges, we experiment with this for regional edge representations of 3x3, 4x4 and 5x5 regions.

Table 14: Results for smoothing the HSV 80x1x1+1 *shot+adj+video* hierarchical colour structure for the *TRECVid 2002* search task.

TRECVID 2002	*Shot+Adj+Video*					*V. Shot Only*	*V. Video Only*	*V. Adj Only*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*
Witten-Bell	**Adj**							
Unbiased	1.00	**.0142**	.034	.024	.018	+1.5% ∼ .172	-15.8% ∼ .604	-16.7% ∼ .285
Best	(10.00	**.0138**	.034	.023	.017)	-1.2% ∼ .825	-18.1% ∼ .587	-42.1% ∼ .944
Jelinek Mercer	$\lambda_{\mathbf{shot}}, \lambda_{\mathbf{adj}}, \lambda_{\mathbf{vid}}, \lambda_{\mathbf{col}}$, **Adj**							
Unbiased	0.95, 0.02, 0.00, 0.02, 1	**.0129**	.026	.024	.018	+3.9% ∼ .407	-21.3% ∼ .598	-18.3% ∼ .324
Best	(0.00, 0.30, 0.60, 0.10, 10	**.0261**	.026	.027	.021)	+109.7% ∼ .839	+56.3% ∼ .901	+14.8% ∼ .458
Absolute	δ							
Unbiased	0.10	**.0140**	.037	.023	.017	-1.2% ∼ .961	-16.9% ∼ .606	-41.3% ∼ .937
Best	0.10	**.0140**	.037	.023	.017	-1.2% ∼ .961	-16.9% ∼ .606	-41.3% ∼ .937
Average of unbiased		**.0137**	.032	.024	.017	+1.4%	-18.0%	-25.4%
Average of best		**.0180**	.032	.025	.019	+35.7%	+7.1%	-22.9%

6 4 1 Global Canny Edge

We evaluated the *Canny 4+1*, *Canny 16+1*, *Canny32+1* and *Canny64+1* visual languages The results for visual language models and standard visual retrieval models for global Canny edge representations are shown in Figure 40 for the TRECVid 2002 and TRECVid 2003 search tasks For both collections the performance in general increases with the number of quantisation levels with *Canny 64+1* producing the best results The results for edges are considerably worse than previously presented for colour The retrieval models' MAP are at best a third of the *HSV 16x4x4* colour results for TRECVid 2002 and for TRECVid 2003 they achieve at best only a half of the *HSV 16x4x4* colours' results For both collections, the language models for the global Canny edge feature perform as well as, or better than, the standard visual retrieval models

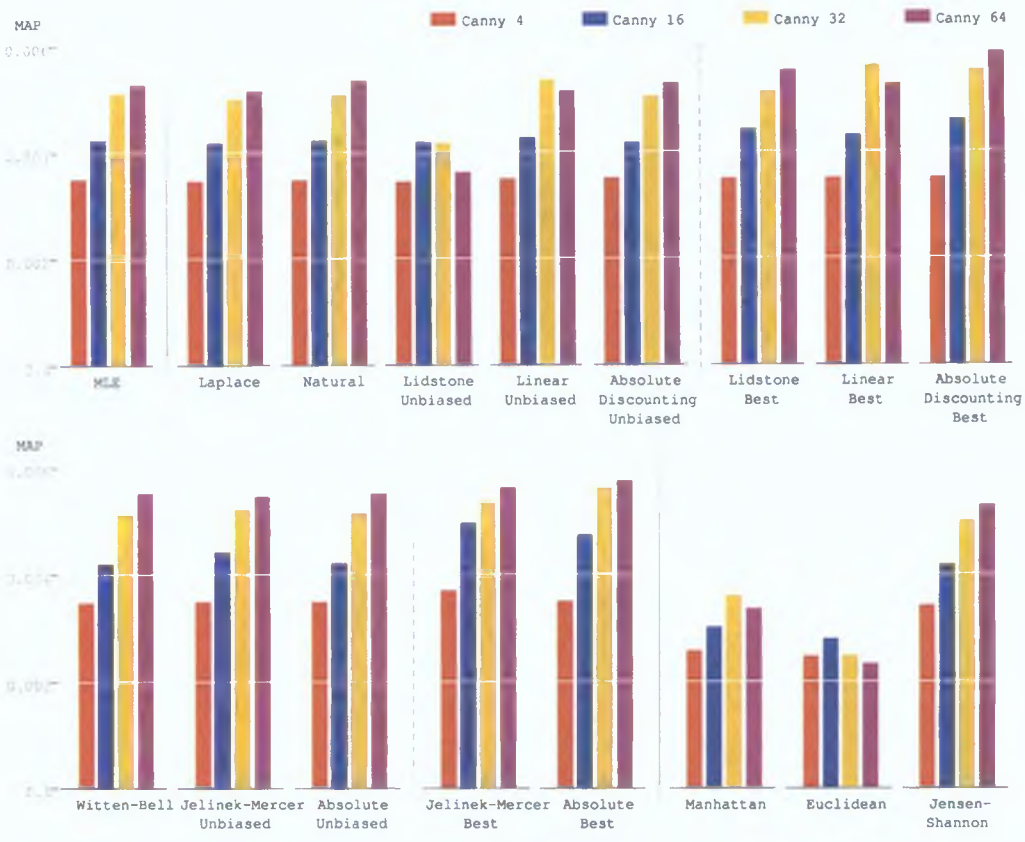
Jensen-Shannon distance is the best performing *standard visual retrieval model* for all tested Canny representations on both collections For the *Canny 64+1* representation, see Tables 47 (appendix page 252) and 48 (appendix page 253), it achieves a MAP of 0 0053 on the TRECVid 2002 collection and a MAP of 0 0083 on the TRECVid 2003 collection The other standard visual models produce similar results to each other, but as the number of quantisation levels increases, see Figure 40, Euclidean distance becomes relatively worse than Manhattan distance

For most Canny representations on both collections, see Figure 40, the MLE language model and the Jensen-Shannon distance produce the same results For the best representation, *Canny 64+1*, the MLE language model has the same performance as Jensen-Shannon distance on TRECVid 2002 and a 9 6% relatively worse MAP for TRECVid 2003

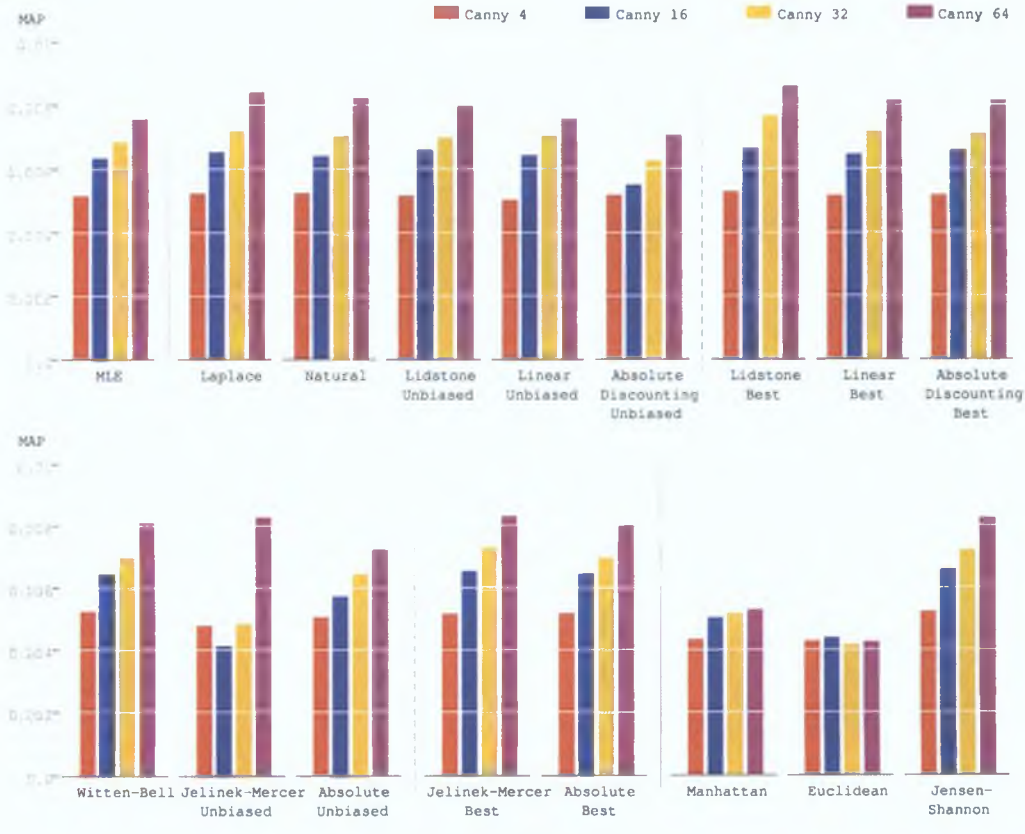
The results for TRECVid 2002 in Figure 40(a) show that only small improvements relative to the MLE model are achieved with smoothing for the global Canny feature on this collection

For the TRECVid 2003 search task Lidstone discounting is the *best optimised language model* for all tested representations with a peak MAP of 0 0086 for *Canny 64+1* which is 7% better than the Jensen-Shannon distance result Even though the optimised parameters for Lidstone discounting on both collections are very different, the unbiased results for Lidstone discounting for the tested Canny representations on TRECVid 2003 never underperform the MLE language model Jelinek-Mercer smoothing is the *best interpolation based language model* on TRECVid 2003 and achieves the same result as Jensen-Shannon distance, but its unbiased results are unstable, performing worse than the MLE model on all but the *Canny 64+1* representation As for global HSV colour, we get some indication from our results for global Canny edges that the visual language models smoothed with a background collection model do not necessarily produce superior results when compared to the discounted language models The performance differences are minuscule for the retrieval models on the global Canny feature so it is difficult to extrapolate much from these results

In Figures 41 and 42 we compare the global Canny representations for both collections on their 20 best performing topic images using the best optimised retrieval model for both collections - Absolute discounting for TRECVid 2002 and Lidstone discounting for TRECVid 2003 For the TRECVid 2002 search task *Canny 64+1* is the best representation for 11 out of the top 20 topic images and *Canny 32+1* is the best representation for 5 of these topic images For TRECVid 2003's top 20 topic images, see Figure 42, *Canny 64+1* and *Canny*



(a) TRECVid 2002



(b) TRECVid 2003

Figure 40: Comparison of *global Canny edge representations* with different number of bins (4, 16, 32, 64) using language models and standard visual retrieval models for the *TRECVid 2002* search task.

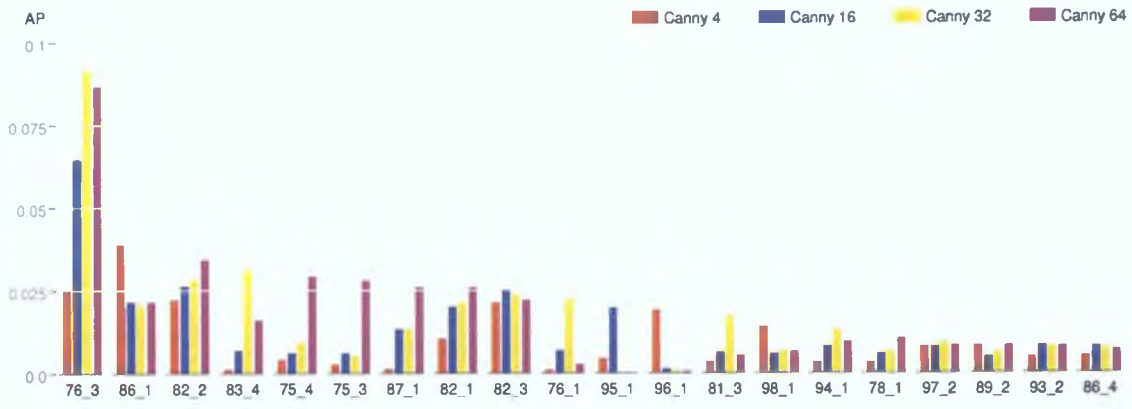


Figure 41: Comparison of *global Canny edge representations* with different number of bins (4, 16, 32, 64) *on the 20 most successful TRECVid 2002 topic images* using the optimised Absolute discounting language model.

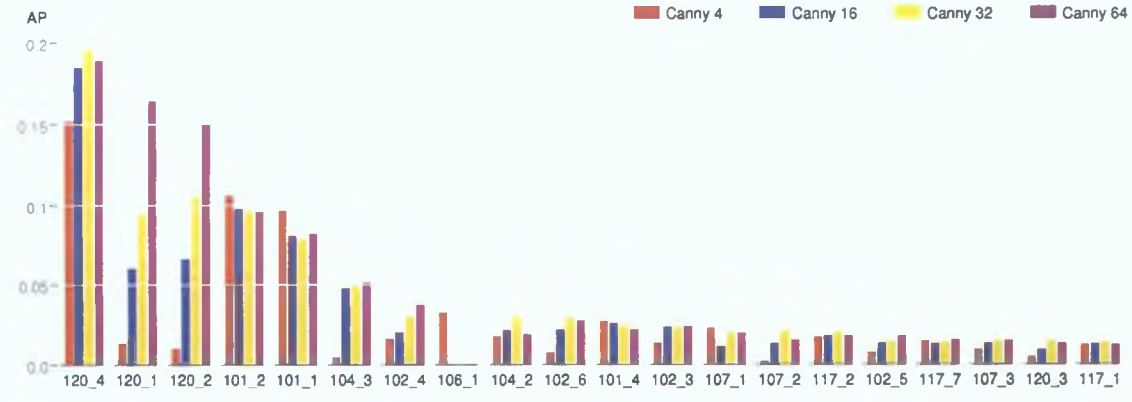


Figure 42: Comparison of *global Canny edge representations* with different number of bins (4, 16, 32, 64) *on the 20 most successful TRECVid 2003 topic images* using the optimised Lidstone language model.

32+1 produce similar results with *Canny 64+1* producing the best average precision for 8 topic images and *Canny 32+1* achieving the best for 7 topic images For the 5 other topic images the compact *Canny 4+1* representation is best Overall for both collections the performance boost for the *Canny 64+1* representation is more evenly spread across the high performing topic images than for the previously discussed structural smoothing and unlike structural smoothing the improvements for *Canny 64+1* occur for many of the official TRECVid topics and not just a couple of images belonging to the same two topic

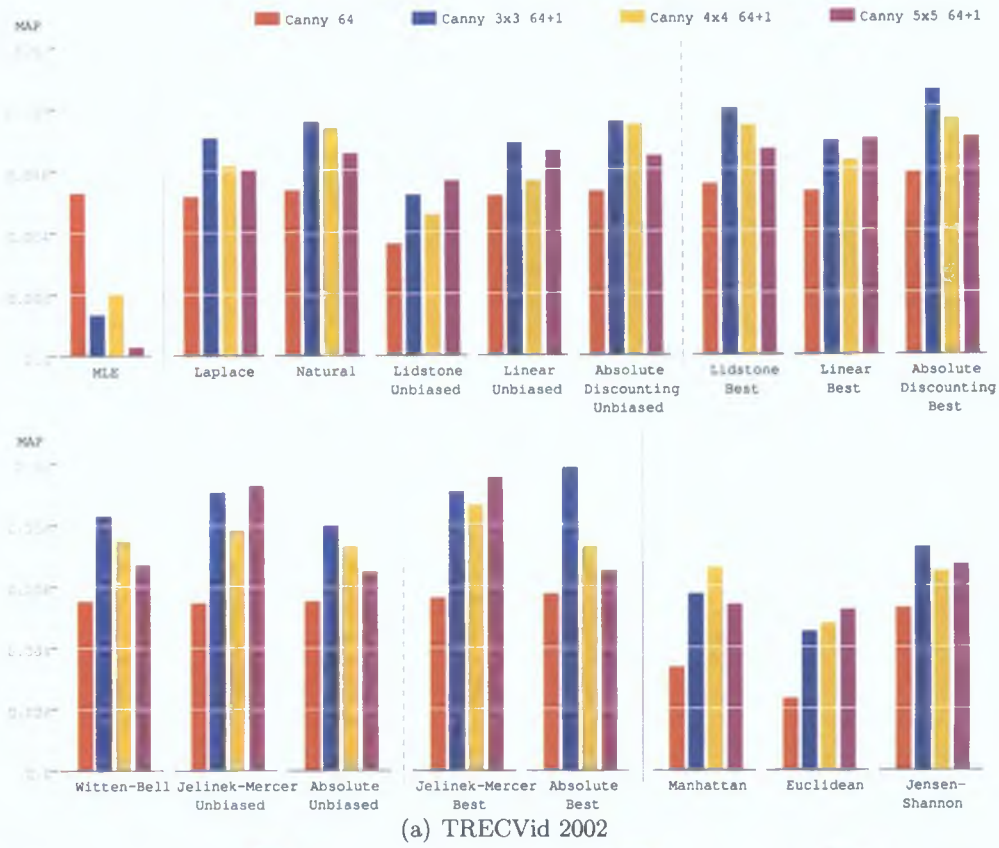
We see from looking at the Wilcoxon p-values in Tables 47 (appendix page 252) and 48 (appendix page 253), which compare *Canny 64+1* with the other Canny representations, that the improvements when using the *Canny 64+1* representation are statistically significant for most of the retrieval models However, for TRECVid 2002 there is not much difference between *Canny 32+1* and *Canny 64+1* with some retrieval models, see Table 47, producing poorer results for the *Canny 64+1* representation and some of the small improvements for the *Canny 64+1* representation are statistically insignificant In contrast, for all the retrieval models for the TRECVid 2003 search task, see Table 48, there is improvement in the MAP measure with many of the models having more than a 10% statistically significant improvement for the *Canny 64+1* representation compared to the *Canny 32+1* representation For this reason we will use the *Canny 64+1* representation in our regional experiments

6 4 2 Regional Canny Edge

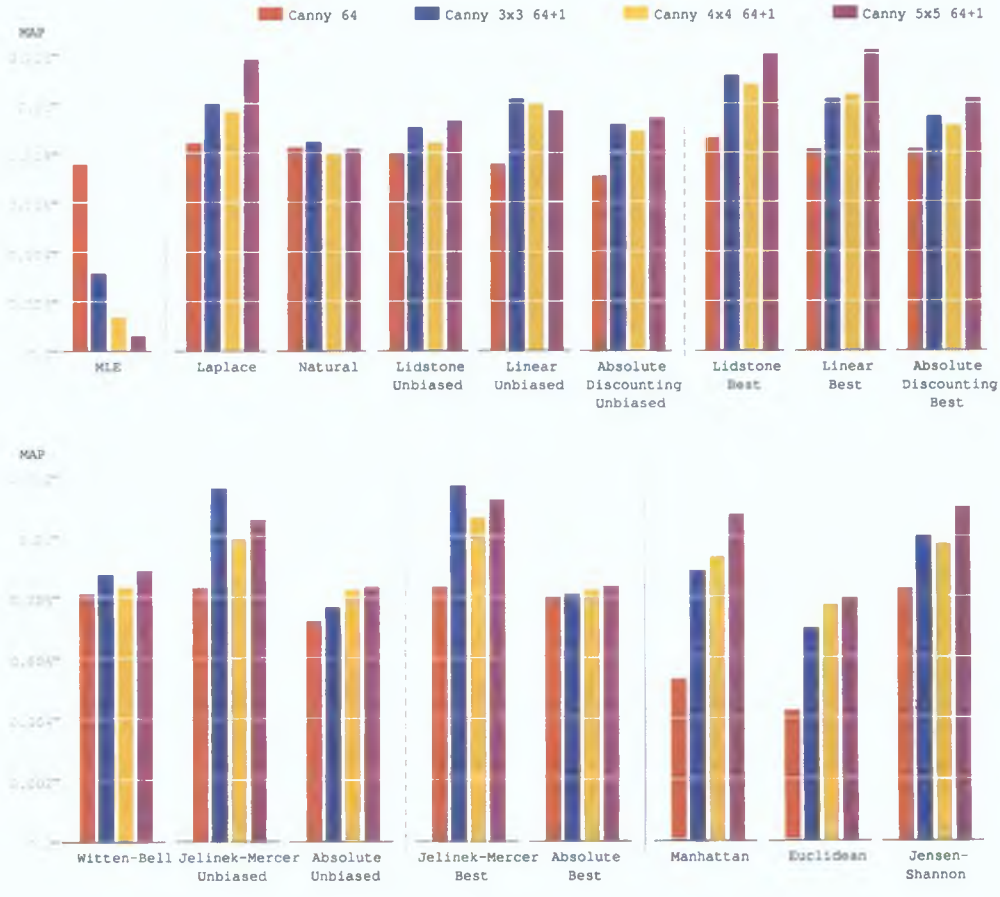
In this section we compare different regional *Canny 64+1* edge representations for 3x3, 4x4 and 5x5 regions These representations produce visual languages consisting of 585, 1040, and 1625 symbols The results are shown in Figure 43 for the TRECVid 2002 and TRECVid 2003 visual search tasks For both collections the regional Canny representations perform better than global Canny in terms of the MAP measure For TRECVid 2002 3x3 regions in general performs best and a slight decrease in performance is observed for more regions For TRECVid 2003, see Figure 43(b), the best performance for most retrieval models is achieved for 5x5 regions

In Figures 44 and 45 we present the results for comparing regional *Canny 64+1* representations on the 20 most successful topic images for the best optimised retrieval models on both collections, Absolute interpolation language model (best MAP of 0 0098 for 3x3 regions) for TRECVid 2002 and the Linear discounting language model (best MAP of 0 0121 for 5x5 regions) for TRECVid 2003 For the Absolute interpolation language model on TRECVid 2002 the regional representations achieve 12 of the top 20 best results - 6 for 3x3 regions, 3 for 4x4 regions, 3 for 3x3 regions and 8 for no regions For the Linear model on TRECVid 2003 the regional representation achieves 17 of the best 20 results - 5 for 3x3 regions, 3 for 4x4 regions, 9 for 5x5 regions and 3 for no regions While we believe that the regional edge model is better than the global model it is more difficult to decide between the different regional representations

The Tables 49 (appendix page 254) and 50 (appendix page 255) present the results for the *Canny 64+1* representation for 3x3 regions for TRECVid 2002 and for 5x5 regions for TRECVid 2003 Though the use of regions for Canny edges in general does not produce statistically significant better results than global representation on both collections, they still overall have a



(a) TRECVid 2002



(b) TRECVid 2003

Figure 43: Comparison of *regional Canny edge representations* (Canny 64+1 for no regions, 3x3, 4x4, and 5x5 regions) using language models and standard visual retrieval models for the *TRECVid 2002* and *TRECVid 2003* search tasks.

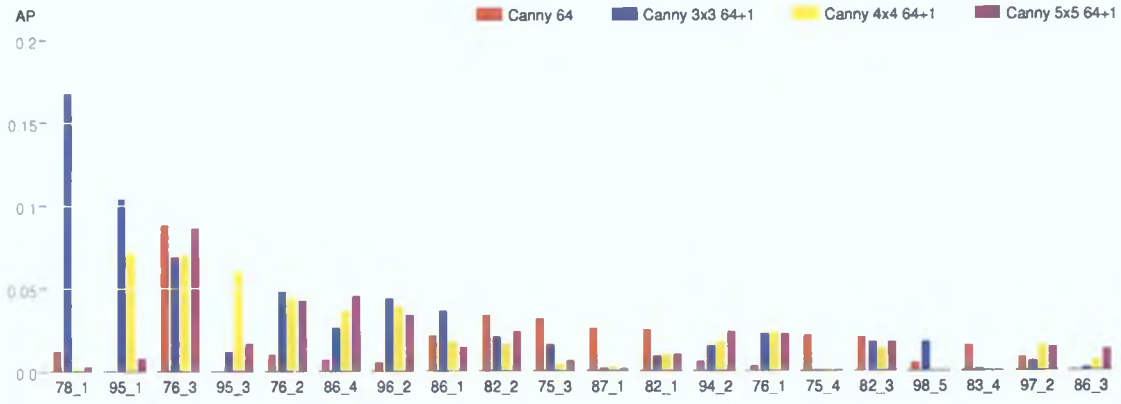


Figure 44: Comparison of *regional Canny edge representations* (Canny 64+1 for no regions, 3x3, 4x4, and 5x5 regions) *on the 20 most successful TRECVid 2002 topic images* using the optimised Absolute interpolation language model.

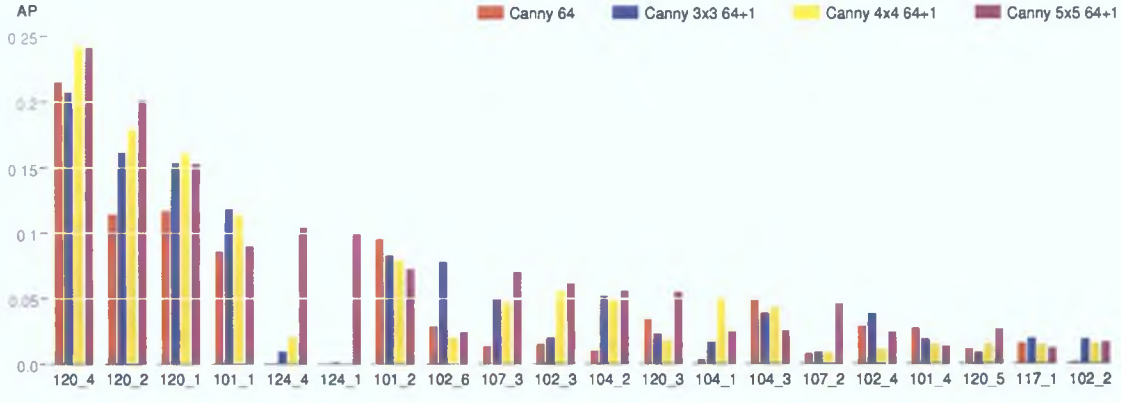


Figure 45: Comparison of *regional Canny edge representations* (Canny 64+1 for no regions, 3x3, 4x4, and 5x5 regions) *on the 20 most successful TRECVid 2003 topic images* using the optimised Linear discounting language model.

positive effect on most of the results For TRECVid 2003 the retrieval models Absolute discounting, Jelinek-Mercer and Manhattan distance achieve statistically significantly better MAP for 5x5 regions than for the global representation according to the Wilcoxon test results and the negative differences between 5x5 regions and the other regional representations are not statistically significant for any of the retrieval models For this reason we will specifically investigate the 5x5 regional representation though implementation criteria such as smaller memory footprint or quicker retrieval time might in practice lead us to choose the 3x3 regional representation since its index size and retrieval speed is a third of that required for 5x5 regions

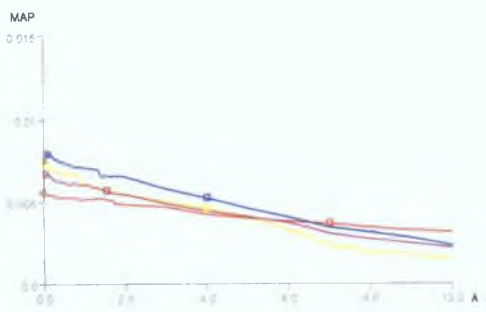
Jensen-Shannon distance is the best standard visual model for both collections For TRECVid 2002 it has its best performance for 3x3 regions with a MAP of 0 0072, a 36% improvement on its global MAP of 0 0053 For TRECVid 2003 it peaks at 5x5 regions with a MAP of 0 0110, a 33% improvement compared to its global MAP of 0 0083 The relative ordering of standard visual models remains the same Jensen-Shannon distance is followed by Manhattan distance with Euclidean distance the worst

The MLE language model, as expected, is useless when using regions due to the increase in vocabulary size Unlike for global edges, smoothing is essential for the regional feature in order for the language modelling approach to achieve similar or better results than the standard visual models

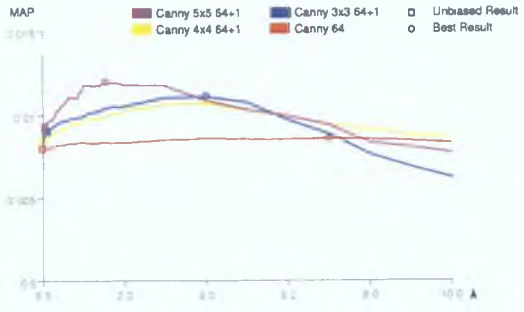
Jelinek-Mercer smoothing achieves consistently good results for both collections for the regional Canny feature On TRECVid 2002 it peaks at 5x5 regions with an optimised MAP of 0 0095, an improvement of 69% on its global MAP of 0 0056, and an unbiased MAP of 0 0092 Interestingly, the amount of smoothing is high with $\lambda = 0\ 85$ for the optimised Jelinek-Mercer model, which means that 85% of the probability mass of a document's smoothed probability model is due to the background collection model Absolute discounting is the best performing discounting language model for TRECVid 2002 and peaks at a MAP of 0 0085 (0 0075 unbiased) for 3x3 regions

For TRECVid 2003 Lidstone discounting is the best discounting language model for most regions except by an insignificant amount for 5x5 regions where it has a MAP of 0 0120 compared to a MAP of 0 0121 for Linear discounting Unfortunately, the optimum parameter for Lidstone smoothing is inconsistent for both collections, for 5x5 regions $\lambda = 0\ 06$ for TRECVid 2002 and $\lambda = 1\ 55$ for TRECVid 2003, producing disappointing unbiased results (see Figures 46(a) and 46(b)) Jelinek-Mercer is the best interpolation based language model for all regions on TRECVid 2003 Though its optimum MAP of 0 0112 is lower than Lidstone, the unbiased results for Jehnek-Mercer perform a lot better and more consistently than for Lidstone smoothing

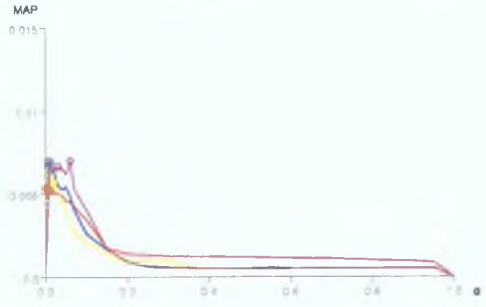
We present statistical significance comparisons of the unbiased retrieval models (except the poor MLE and Euclidean distance models) on the *Canny 5x5 64+1* feature for the set of topic results across the three TRECVid collections in Table 15 This table indicates that the Jelinek-Mercer language model is the best retrieval model for this regional Canny feature The Jensen-Shannon distance is also very strong and is second in the rankings in terms of MAP but is not statistically significantly poorer than the Jelinek-Mercer language model Both the Jelinek-Mercer language model and Jensen-Shannon distance are statistically significantly better than



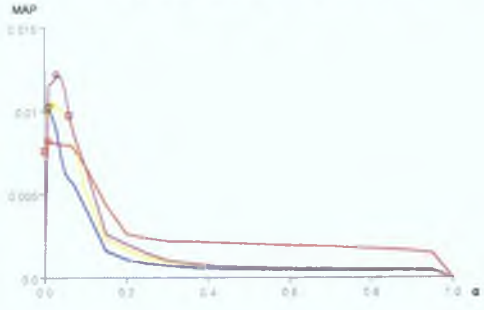
(a) TRECVid 2002: Lidstone



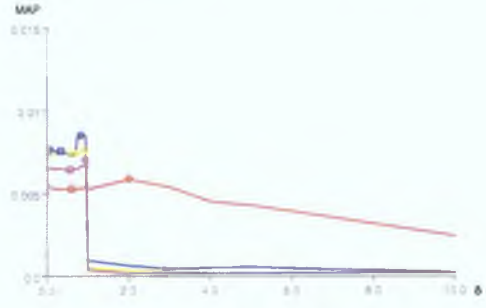
(b) TRECVid 2003: Lidstone



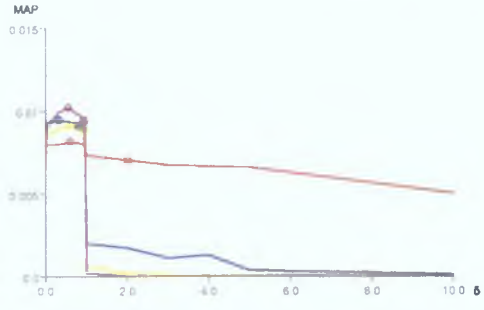
(c) TRECVid 2002: Linear



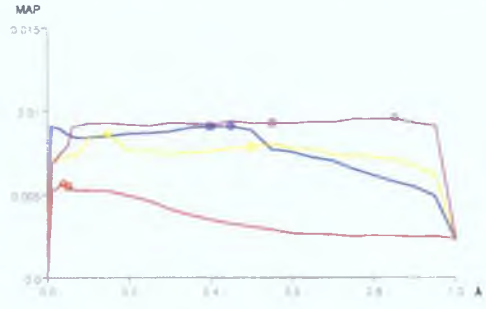
(d) TRECVid 2003: Linear



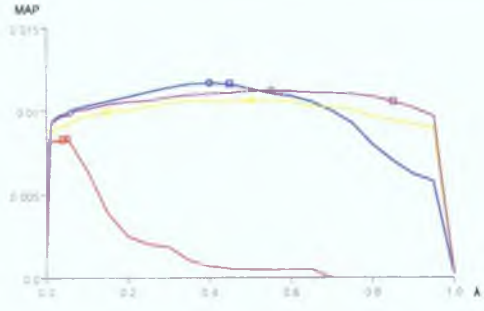
(e) TRECVid 2002: Absolute Discounting



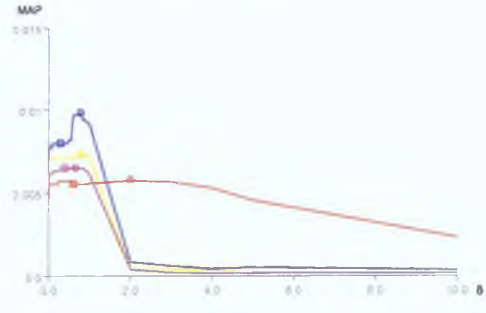
(f) TRECVid 2003: Absolute Discounting



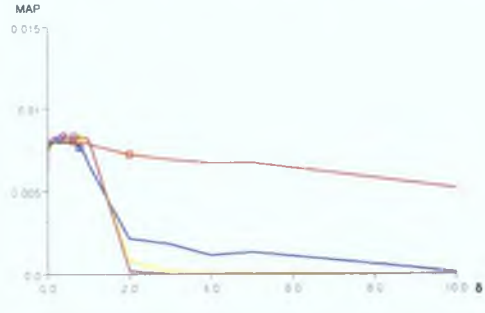
(g) TRECVid 2002: Jelinek-Mercer



(h) TRECVid 2003: Jelinek-Mercer



(i) TRECVid 2002: Absolute Interpolation



(j) TRECVid 2003: Absolute Interpolation

Figure 46: Plot of MAP over the parameter space for the parametric language models using regional Canny edges for *TRECVid 2002* search task and *TRECVid 2003* search task.

Table 15: Statistical significance comparison of retrieval models on the regional 5x5 Canny 64+1 edge feature for the aggregated TRECVid 2002, 2003 and 2004 search tasks.

Ret. Meth.	*MAP/%Dif*	*Wilcoxon Test Results*								
Jelinek-Mercer (JM)		-	>JS	>Man	>Lap	>AbsD	>WB	>Lin	>Abs	>Lid
(JM)	0.0092		.149	.015	.036	.166	.102	.003	.023	.009
Jensen-Shannon (JS)		<JM	-	>Man	>Lap	>AbsD	>WB	>Lin	>Abs	>Lid
(JS)	0.0079/-14.1%	.149		.000	.001	.495	.218	.001	.038	.000
Manhattan (Man)		<JM	<JS	-	>Lap	>AbsD	>WB	>Lin	>Abs	>Lid
(Man)	0.0077/-16.1%	.015	.000		.788	.965	.602	.092	.355	.115
Laplace (Lap)		<JM	<JS	<Man	-	>AbsD	>WB	>Lin	>Abs	>Lid
(Lap)	0.0075/-18.5%	.036	.001	.788		.883	.572	.005	.238	.000
Absolute Disc. (AbsD)		<JM	<JS	<Man	<Lap	-	>WB	>Lin	>Abs	>Lid
(AbsD)	0.0072/-21.6%	.166	.495	.965	.883		.030	.004	.000	.000
Witten-Bell (WB)		<JM	<JS	<Man	<Lap	<AbsD	-	>Lin	>Abs	>Lid
(WB)	0.0067/-26.7%	.102	.218	.602	.572	.030		.086	.008	.214
Linear (Lin)		<JM	<JS	<Man	<Lap	<AbsD	<WB	-	>Abs	>Lid
(Lin)	0.0066/-28.3%	.003	.001	.092	.005	.004	.086		.615	.376
Absolute (Abs)		<JM	<JS	<Man	<Lap	<AbsD	<WB	<Lin	-	>Lid
(Abs)	0.0063/-31.4%	.023	.038	.355	.238	.000	.008	.615		.693
Lidstone (Lid)		<JM	<JS	<Man	<Lap	<AbsD	<WB	<Lin	<Abs	-
(Lid)	0.0063/-31.6%	.009	.000	.115	.000	.000	.214	.376	.693	

all other retrieval models except Absolute discounting and Witten-Bell smoothed language models. Lidstone smoothing, which is equivalent to interpolation with an uniform source in our experiments, performs worst with 31% less MAP on average across the three TRECVid collections in comparison to Jelinek-Mercer smoothing. The background collection model seems to be important for this regional feature probably because it has a skewed distribution due to the regional non-edge histogram bins containing most of the probability mass.

We tabulate the results of our tested retrieval models on the official TRECVid multi-example topics in Table 16 that fused the single image results using CombSUM. The results are about 50% lower for TRECVid 2002 and 2004 and about 100% lower for TRECVid 2003 compared to our regional colour results. The results for the Jelinek-Mercer language model are 50.5% and 57.1% less than for the DCT GMM model for TRECVid 2002 and 2003, and 155.0% better than the DCT GMM model for TRECVid 2004. The mean result for the Jelinek-Mercer language model on TRECVid 2004 is possibly an outlier since it is largely due to excellent results on two topics (topic 135 and 142). Overall the results are encouraging for this limited feature and given that it contains no colour information, it should provide additional information when fused with the colour results.

6.5 Experiments with Texture Feature

In this section we investigate language models and visual retrieval models using texture features for the TRECVid query-by-example task. In our experiments we represent texture by using different numbers of significant DCT coefficients into a fixed number of quantisation levels, as described in Chapter 4. As before, we will first explore texture as a global feature before

Table 16: Comparison of retrieval models on the Canny 5x5 64+1 feature for official TRECVid topics (i.e. fused topic examples).

VisExsEdge-CombScore	*TRECVid 2002*				*TRECVid 2003*				*TRECVid 2004*				*TRECVid 02-04*			
Retrieval Method	*MAP*	*P10*	*P30*	*P100*	*MAP*	*P10*	*P30*	*P100*	*MAP*	*P10*	*P30*	*P100*	*MAP*	*P10*	*P30*	*P100*
Laplace	**.0103**	.052	.028	.016	**.0190**	.064	.057	.033	**.0087**	.035	.029	.019	**.0128**	.051	.038	.022
Best	(**.0109**	.052	.029	.015)	(**.0192**	.064	.057	.032)	(**.0088**	.035	.030	.019)	(**.0131**	.051	.039	.022)
Lidstone	**.0067**	.040	.019	.013	**.0165**	.064	.057	.031	**.0069**	.026	.016	.015	**.0101**	.044	.031	.020
Best	(**.0117**	.068	.037	.019)	(**.0196**	.060	.049	.033)	(**.0073**	.022	.017	.013)	(**.0130**	.051	.035	.022)
Linear	**.0061**	.024	.023	.012	**.0141**	.048	.039	.022	**.0064**	.017	.017	.011	**.0089**	.030	.026	.015
Best	(**.0078**	.036	.020	.012)	(**.0187**	.064	.047	.035)	(**.0071**	.017	.017	.011)	(**.0113**	.040	.028	.020)
Absolute Discounting	**.0109**	.056	.028	.022	**.0148**	.076	.061	.033	**.0106**	.048	.036	.021	**.0121**	.060	.042	.026
Best	(**.0127**	.064	.033	.020)	(**.0168**	.076	.065	.038)	(**.0106**	.048	.036	.019)	(**.0134**	.063	.045	.026)
Witten-Bell	**.0115**	.072	.033	.020	**.0146**	.060	.055	.031	**.0083**	.035	.029	.017	**.0115**	.056	.039	.023
Best	(**.0118**	.072	.036	.020)	(**.0148**	.060	.055	.032)	(**.0083**	.035	.029	.018)	(**.0117**	.056	.040	.023)
Jelinek-Mercer	**.0142**	.072	.033	.020	**.0133**	.048	.037	.022	**.0255**	.078	.054	.027	**.0174**	.066	.041	.023
Best	(**.0146**	.076	.036	.021)	(**.0183**	.056	.051	.032)	(**.0255**	.078	.054	.027)	(**.0193**	.070	.047	.027)
Absolute	**.0105**	.056	.029	.019	**.0129**	.060	.048	.027	**.0060**	.030	.029	.015	**.0099**	.049	.036	.021
Best	(**.0117**	.064	.033	.018)	(**.0133**	.064	.052	.028)	(**.0078**	.035	.029	.016)	(**.0110**	.055	.038	.021)
Manhattan	**.0096**	.056	.024	.014	**.0187**	.064	.056	.028	**.0183**	.065	.042	.025	**.0154**	.062	.041	.022
Best	(**.0107**	.060	.029	.015)	(**.0199**	.056	.053	.032)	(**.0183**	.065	.042	.025)	(**.0162**	.060	.042	.024)
Jensen Shannon	**.0101**	.052	.027	.017	**.0194**	.064	.059	.032	**.0120**	.039	.041	.020	**.0139**	.052	.042	.023
Best	(**.0124**	.060	.031	.019)	(**.0199**	.068	.052	.037)	(**.0120**	.039	.039	.022)	(**.0148**	.056	.041	.026)
Euclidean	**.0094**	.044	.023	.010	**.0139**	.064	.048	.025	**.0096**	.057	.032	.018	**.0110**	.055	.034	.018
Best	(**.0105**	.056	.023	.014)	(**.0143**	.052	.047	.026)	(**.0096**	.057	.032	.018)	(**.0115**	.055	.034	.019)
Average of unbiased	**.0099**	.052	.027	.017	**.0157**	.061	.052	.028	**.0112**	.043	.032	.019	**.0123**	.052	.037	.021
Average of best	**.0115**	.061	.031	.017	**.0175**	.062	.053	.033	**.0115**	.043	.033	.019	**.0136**	.056	.039	.023

experimenting with it for regional texture representations of 3x3, 4x4 and 5x5 regions

6 5 1 Global Texture

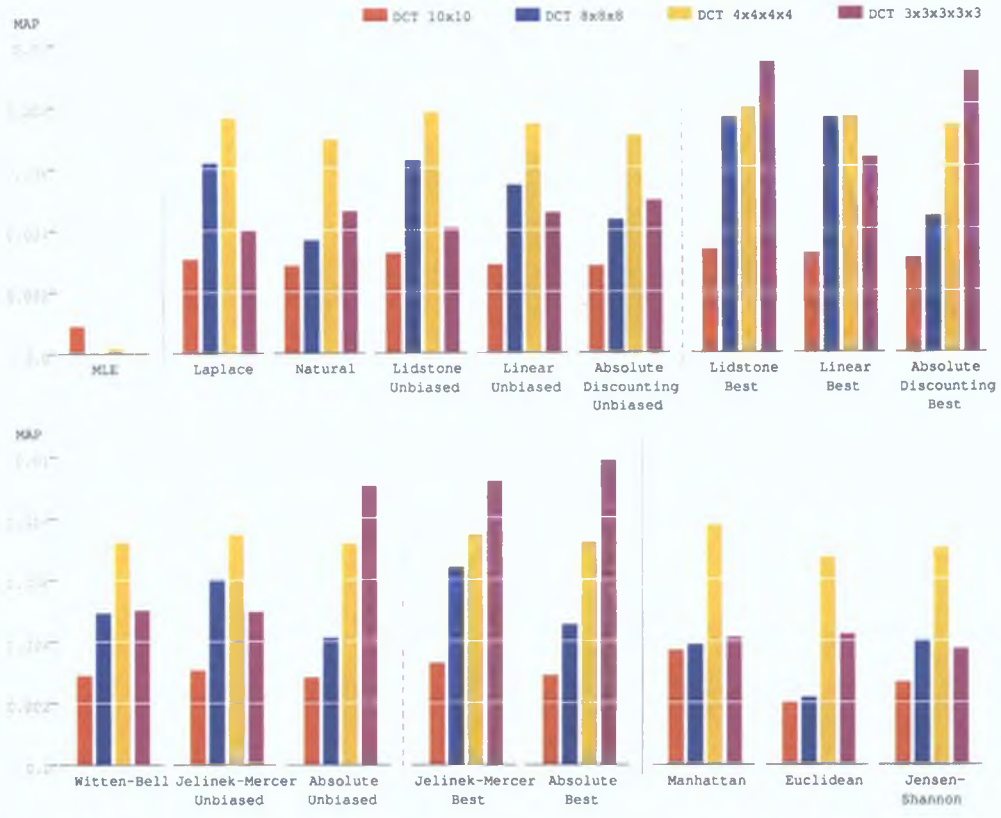
The results for our global DCT texture representations are shown in Figures 47(a) and 50 for the TRECVid 2002 and TRECVid 2003 search tasks We compare texture representations that use between 2 and 5 DCT coefficients Specifically, we compare DCT 10x10, DCT 8x8x8, DCT 4x4x4x4 and DCT 3x3x3x3x3 which produce languages consisting of 100, 256, 256 and 243 symbols respectively

For the TRECVid 2002 collection, see Figure 47(a), DCT 3x3x3x3x3 is the best performing texture representation according to the MAP measure The Absolute interpolation language model achieves the highest MAP for the DCT 3x3x3x3x3 for both the unbiased and optimised results All of the optimised language models except the Linear language model produce their best results on the DCT 3x3x3x3x3 representation, whereas, all the unbiased results except Absolute interpolation produce their best results on the DCT 4x4x4x4 representation These conflicting results from optimised (biased) and unbiased retrieval models can be explained by looking at the individual topic results from the optimised models

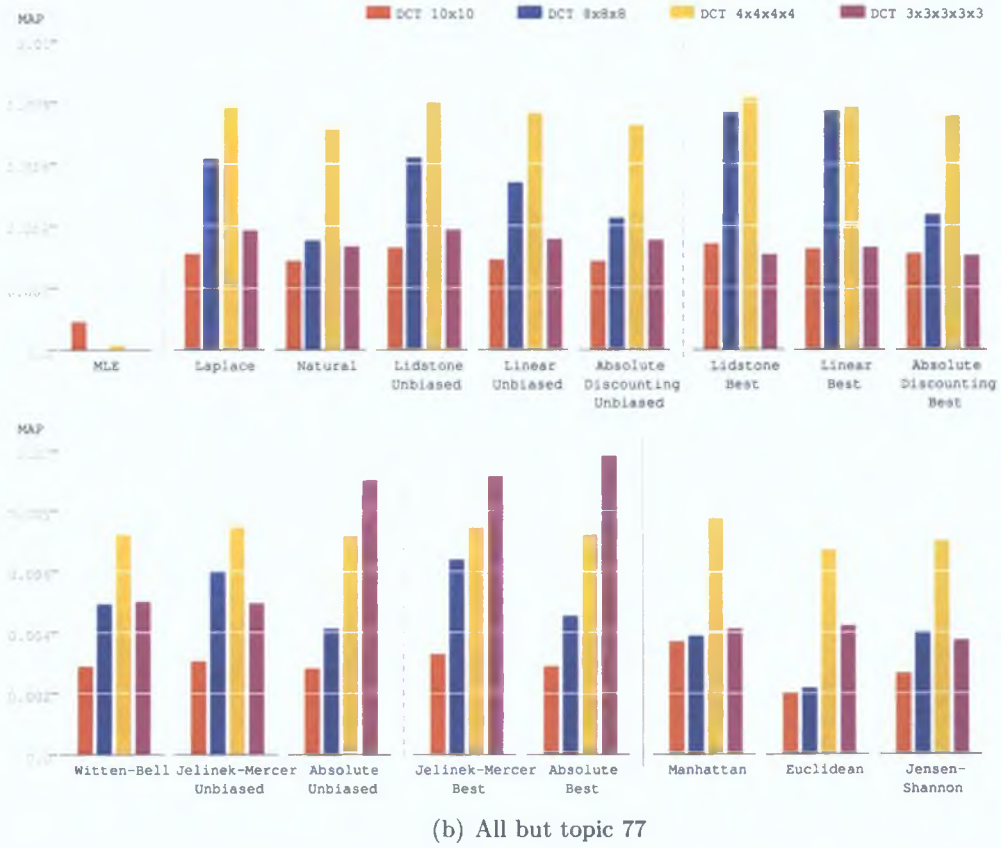
We can see from Figure 48 for the best performing retrieval model on TRECVid 2002, optimised Absolute interpolation, that for its top 20 topic images a single topic image $topic77_1$ (George Washington) accounts for most of the performance gain of the DCT 3x3x3x3x3 representation The DCT 3x3x3x3x3 representation is the best representation for only 3 of the top 20 topic images, while the DCT 4x4x4x4 representation is actually the best performing representation for the majority and produces the best results for 13 of the top 20 topic images The DCT 8x8x8 representation is best for only 3 of the images and the poorly performing DCT 10x10 representation is best for only one topic of the top 20 topic images The results for topic 77 can dwarf all other topic results in the calculation of the MAP statistic since it has only 3 relevant shots and when one or two of these occur in high ranking positions an extremely high average precision in comparison to the other topic results is achieved In the case of the optimised language models and the unbiased Absolute interpolation language model the results for DCT 3x3x3x3x3 have 2 of the 3 relevant documents in the top five ranked documents, which produces an average precision so high compared to all the other topics that it dominates the MAP statistic at the expense of the relative performance of the other topics

Overall, we believe that the DCT 4x4x4x4 representation is the best tested texture representation for the TRECVid 2002 search tasks We can support this claim with the retrieval results for the TRECVid 2002 collection, shown in Figure 47(b), which exclude topic 77 This figure provides a more consistent view of the relative performance of the four DCT texture representations The DCT 4x4x4x4 representation is now clearly the best representation for all the retrieval models for both the unbiased and optimised results The DCT 8x8x8 representation is the next best followed by the DCT 3x3x3x3x3 representation, which now produces nearly similar results to the poorly performing DCT 10x10 representation

In Table 51 (appendix page 256) we tabulate the results for the DCT 4x4x4x4 representation compared with the DCT 3x3x3x3x3 and DCT 8x8x8 representations for the complete TRECVid



(a) All topics



(b) All but topic 77

Figure 47: Comparison of *global DCT representations* (DCT 10x10, DCT 8x8x8, DCT 4x4x4x4, and DCT 3x3x3x3x3) using language models and standard visual retrieval models for the *TRECVid 2002* search task over (a) all topics and (b) all but topic 77.

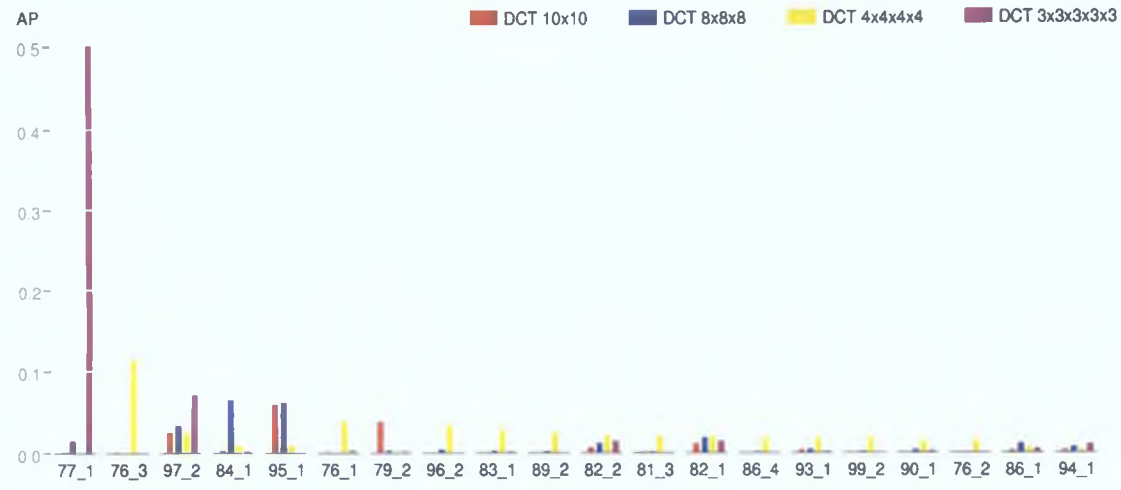
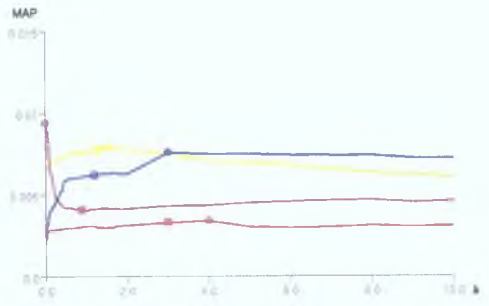


Figure 48: Comparison of *global DCT representations* (DCT 10x10, DCT 8x8x8, DCT 4x4x4x4 and DCT 3x3x3x3x3) *on the 20 most successful TRECVid 2002 topic images* using the optimised Absolute interpolation language model.

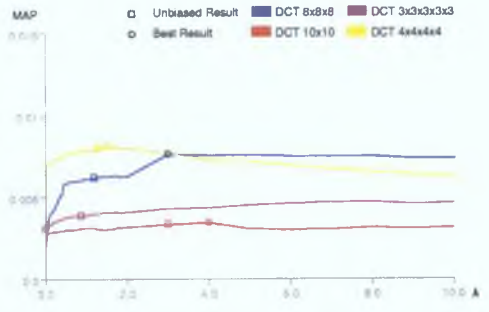
2002 search task. The DCT 4x4x4x4 representation is better than the DCT 8x8x8 for all retrieval models and importantly most of these improvements are statistically significant. For all the unbiased results except the very poor MLE model, DCT 4x4x4x4 is also significantly better than the DCT 3x3x3x3x3 representation. The exceptions are Absolute interpolation and all the optimised language models, except the Linear discounting language model, which produce their best results for the DCT 3x3x3x3x3 representation. However these improvements for the DCT 3x3x3x3x3 representation relative to the DCT 4x4x4x4 representation are not statistically significantly according to the Wilcoxon sign rank test. As can be seen from Table 51 they achieve a p-value of close to 1.0, which is as insignificant as you can get with this statistical test. This further supports our claim that the DCT 4x4x4x4 representation is superior to the DCT 3x3x3x3x3 representation for the TRECVid 2002 collection even though the MAP measure indicates otherwise.

The results for the parametric language models are shown in Figure 49 for the active range of their parameter values for the global DCT texture representations. The results for Lidstone, Absolute discounting and Jelinek-Mercer show that even though DCT 3x3x3x3x3 achieves a higher MAP than the other DCT representations it is only for very specific parameter values. For the majority of parameter values and more importantly for the unbiased results, the MAP for DCT 3x3x3x3x3 is significantly lower than the DCT 4x4x4x4 representation. If we remove topic 77 from the TRECVid 2002 test set, also shown in Figure 49, then the small spikes in the MAP for the DCT 3x3x3x3x3 representation are removed leading to a more consistent view of the MAP response to the parameter values for these language models. Taking this complication into account we will now look more closely at the results for the discounted language models, the interpolated language models and the standard visual retrieval models.

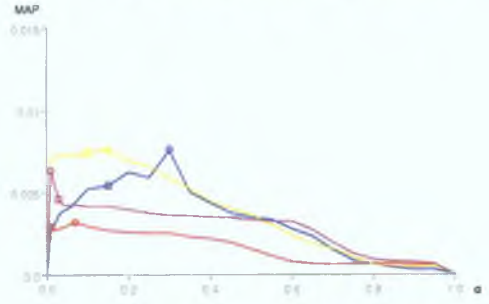
The unsmoothed MLE language model performs very poorly for all DCT representations, while the Lidstone language model, see Figure 47, is the best discounting-based language model for both the unbiased and optimised results even when topic 77 is removed from the test set. The Lidstone model performs best on the DCT 4x4x4x4 representation except for its optimised result on the full TRECVid 2002 topic list in which it favours the DCT 3x3x3x3x3 representation.



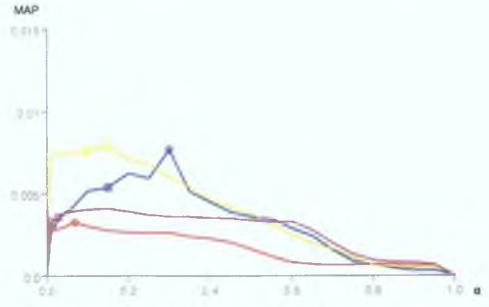
(a) TRECVid 2002: Lidstone



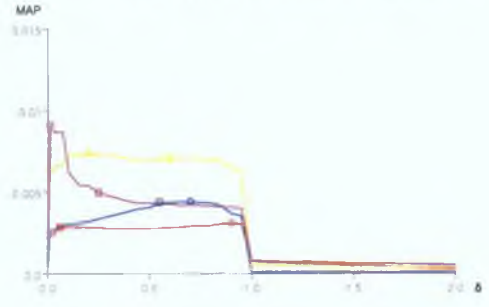
(b) TRECVid 2002 X77: Lidstone



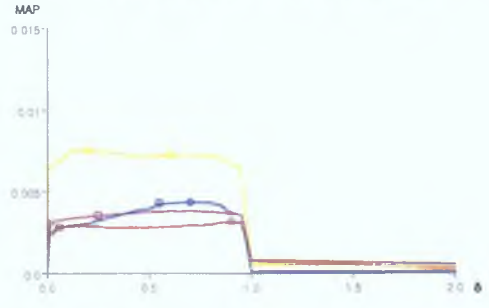
(c) TRECVid 2002: Linear



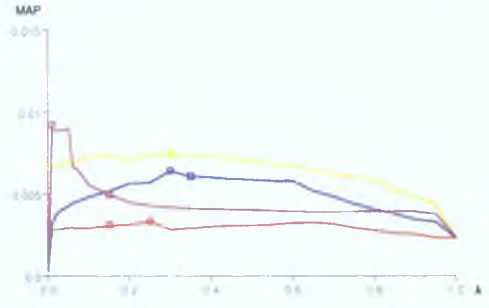
(d) TRECVid 2002 X77: Linear



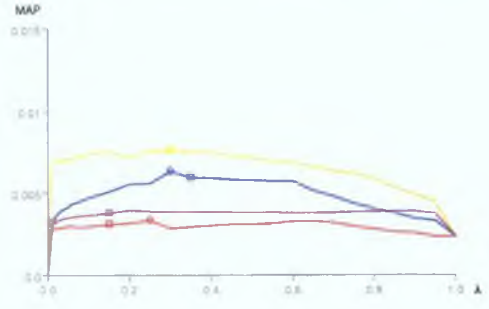
(e) TRECVid 2002: Absolute Discounting



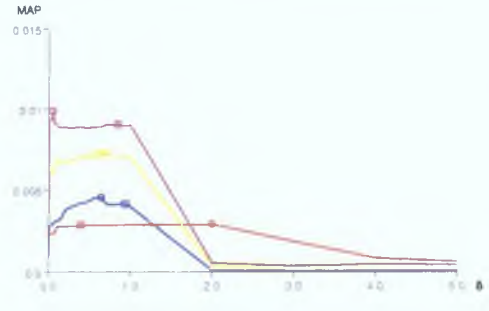
(f) TRECVid 2002 X77: Absolute Discounting



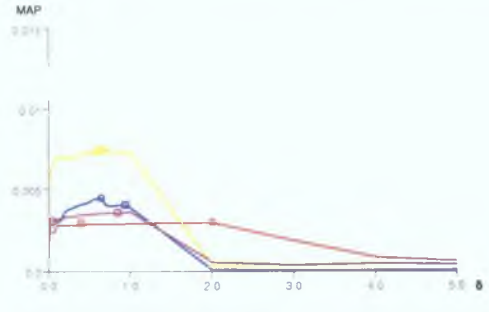
(g) TRECVid 2002: Jelinek-Mercer



(h) TRECVid 2002 X77: Jelinek-Mercer



(i) TRECVid 2002: Absolute Interpolation



(j) TRECVid 2002 X77: Absolute Interpolation

Figure 49: Comparison of MAP for the different parameter values of the parametric language models (Lidstone, Linear, Absolute discounting, Jelinek-Mercer and Absolute interpolation) using the global DCT texture representations (DCT 10x10, DCT 8x8x8, DCT 4x4x4x4 and DCT 3x3x3x3x3) for the *TRECVid 2002* search task for all topics and for all but topic 77.

As can be seen from Figures 49(a) and 49(b) when we exclude topic 77, the performance of the Lidstone discounting language model is stable with respect to changes in its parameter value λ. In this figure the curve between best and unbiased results for all but the DCT 8x8x8 representation is quite flat indicating stability in the choice of the parameter value between the two TRECVid collections.

The Jelinek-Mercer language model is the best collection-based smoothed language model for all but the DCT 3x3x3x3x3 representation in which Absolute interpolation slightly outperforms it. Absolute interpolation is the best retrieval model according to the MAP measure achieving the highest result on the DCT 3x3x3x3x3 representation. The unbiased Absolute interpolation also achieves the highest unbiased result again on the DCT 3x3x3x3x3 representation. However, Absolute interpolation produces poorer results than Jelinek-Mercer smoothing for both its optimised and unbiased results when we remove topic 77 from the test set. We see from Figures 49(g) and 49(h) that topic 77 causes the Jelinek-Mercer language model's MAP to peak for an unusually low amount of smoothing, which is controlled by its λ parameter. When we remove topic 77, the curves for the different DCT representations become more consistent with each other. It can also be observed for Jelinek-Mercer smoothing in Figure 49(g) that the unbiased parameter estimates for λ are very close to the best optimised values for all the DCT representations except of course the problematic DCT 3x3x3x3x3 representation.

In contrast to the previous features, see Figure 47, Manhattan distance is the best standard visual retrieval model. Jensen-Shannon distance is the next best and is followed by the nearly always poorly performing Euclidean distance. All the tested standard visual retrieval models perform best on the DCT 4x4x4x4 representation and when we remove topic 77 from the test set. Manhattan and Lidstone smoothing achieve the joint best result for TRECVid 2002.

The TRECVid 2003 results displayed in Figure 50 show that the DCT 8x8x8 representation is the best tested DCT representation for all retrieval models except for the very poor performing MLE language model. The DCT 3x3x3x3x3 representation is the next best representation and is itself followed by the DCT 10x10 representation. Surprisingly the DCT 4x4x4x4 representation, which was the best representation for the TRECVid 2002 search task when we excluded topic 77, is the worst representation for most of the retrieval models for the TRECVid 2003 collection. The performance of texture is better for the TRECVid 2003 search task than for the TRECVid 2002 search task and the differences between the texture representations are also not as large as on the TRECVid 2003 collection. The higher visual quality of the TRECVid 2003 collection may better support texture features than the TRECVid 2002 collection. Without further research on more video collections, we can only offer this change in visual quality as a probable cause for the differences in the texture representations' performances. The results for the best performing representation DCT 8x8x8 are presented in Table 52 and show that nearly all the unbiased language models perform statistically significantly better for the DCT 8x8x8 representation than for the next best representation DCT 3x3x3x3x3. The larger differences between DCT 8x8x8 and DCT 4x4x4x4 are not statistically significant, though if we look at the topic results we can see that DCT 8x8x8 is a more beneficial representation.

The results for the best retrieval model on TRECVid 2003, the optimised Jelinek-Mercer language model, for the DCT texture feature are shown in Figure 51 for the best 20 topic images. The best representation DCT 8x8x8 produces 9 of the best 20 results, DCT 3x3x3x3x3 produces

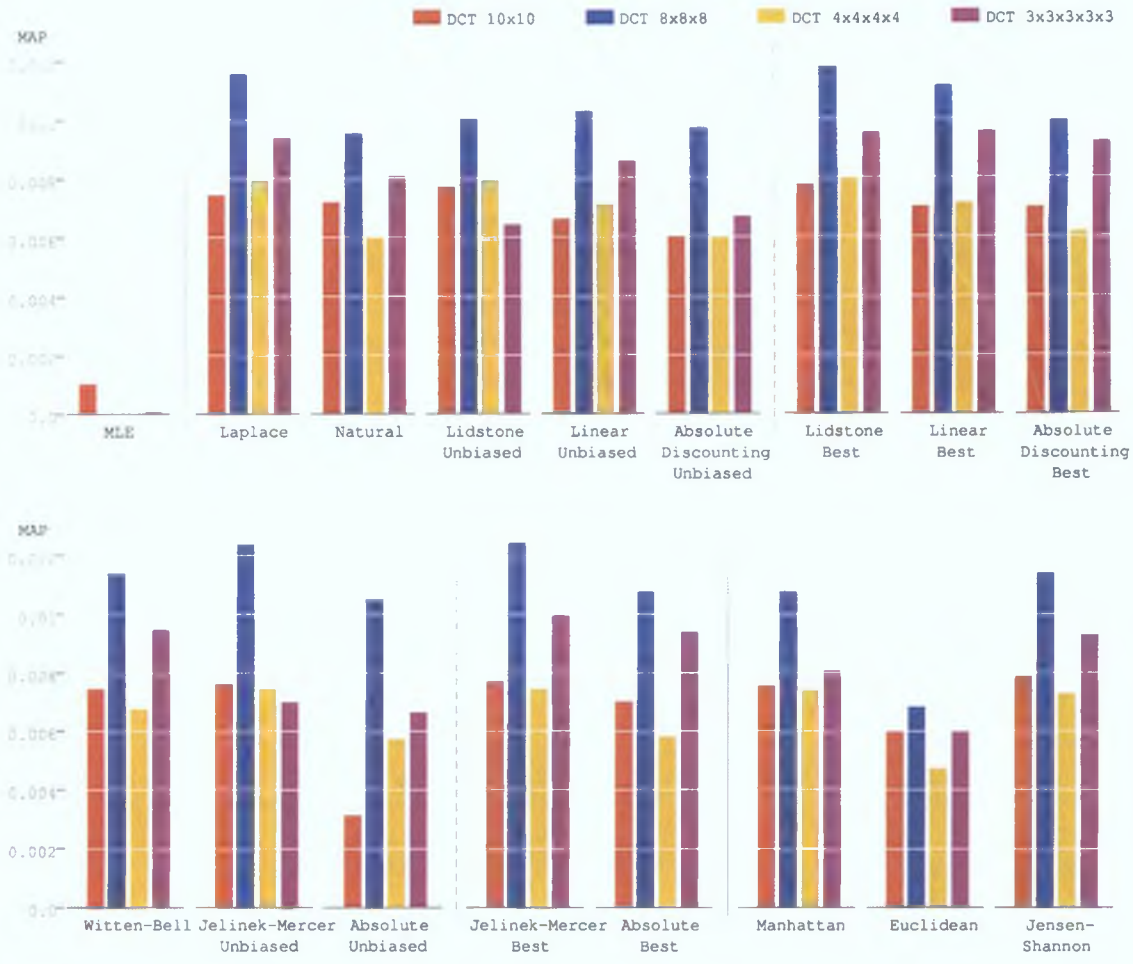


Figure 50: Comparison of *global DCT representations* (DCT 10x10, DCT 8x8x8, DCT 4x4x4x4 and DCT 3x3x3x3x3) using language models and standard visual retrieval models for the *TRECVid 2003* search task.

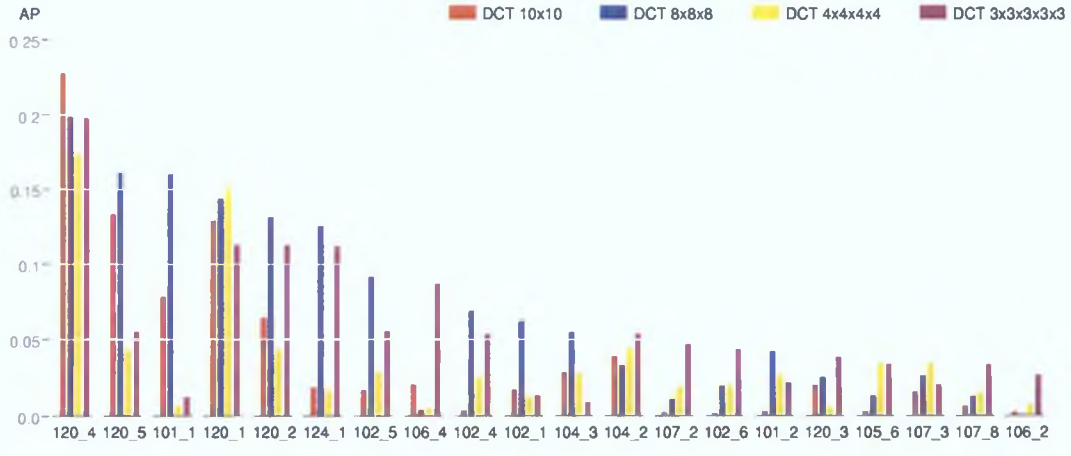
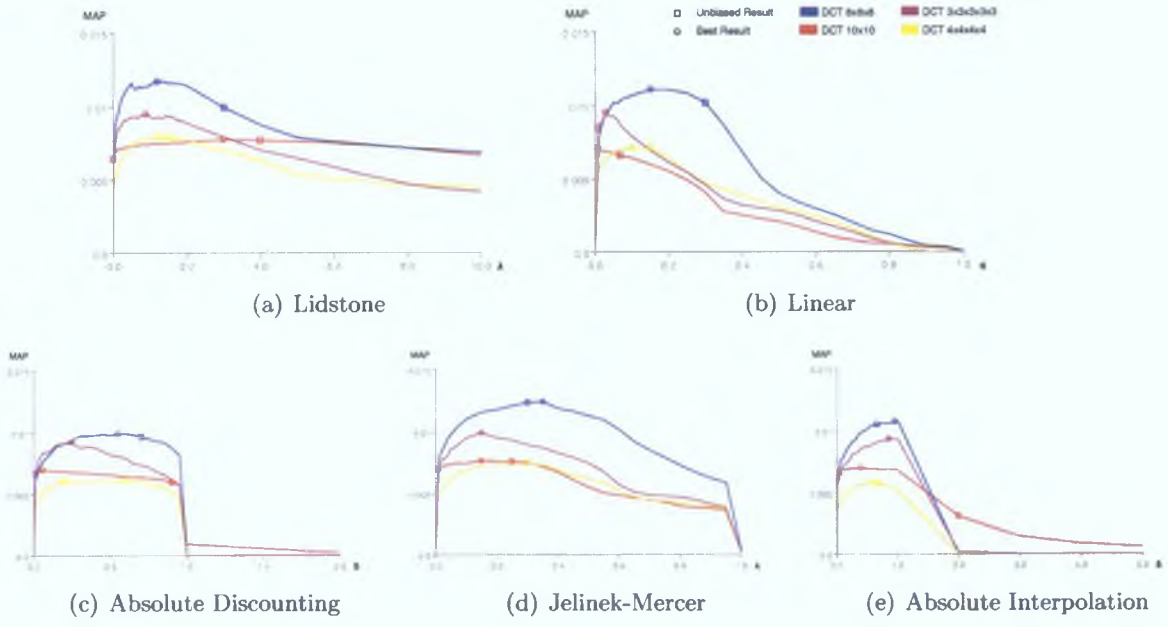


Figure 51: Comparison of *global DCT representations* (DCT 10x10, DCT 8x8x8, DCT 4x4x4x4 and DCT 3x3x3x3x3) *on the 20 most successful TRECVid 2003 topic images* using the optimised Jelinek-Mercer language model.



(a) Lidstone (b) Linear

(c) Absolute Discounting (d) Jelinek-Mercer (e) Absolute Interpolation

Figure 52: Comparison of MAP for the different parameter values of the parametric language models (Lidstone, Linear, Absolute discounting, Jelinek-Mercer and Absolute interpolation) using the global DCT texture representations (DCT 10x10, DCT 8x8x8, DCT 4x4x4x4 and DCT 3x3x3x3x3) for the *TRECVid 2003* search task.

7 of the best 20 results and DCT 4x4x4x4 produces 4 of them. The DCT 10x10 representation only performs best for 1 of the top 20 images. Unlike the results for the best retrieval model for TRECVid 2002 in Figure 48, these results for the TRECVid 2003 topics are not dominated by a single topic image. The MAP statistics for the TRECVid 2003 collection are therefore less skewed than for the TRECVid 2002 collection. For the 20 worst performing topic images DCT 4x4x4x4 produces better results than the DCT 8x8x8 representation and this result goes some way to explaining the lack of statistical significance between their representations for the TRECVid 2003 collection.

For the TRECVid 2003 search task the best unbiased discounting model for all but the DCT 10x10 representation, see Figure 50, is Laplace smoothing. Laplace smoothing is of course equivalent to Lidstone smoothing when its parameter λ is set to 1 and Lidstone smoothing is in fact the best unbiased discounting model for the DCT 10x10 representation. Similar to TRECVid 2002 the best optimised discounting method is Lidstone discounting which peaks, see Table 52, at an optimised MAP of 0.0118 for the DCT 8x8x8 representation. The optimised Lidstone discounting is the best discounting method for all tested representations except insignificantly with the DCT 3x3x3x3x3 representation, see Table 53, in which Linear achieves a MAP of 0.0096 compared to Lidstone's MAP of 0.0095.

The unbiased results for the Lidstone language model are reasonable except for the DCT 3x3x3x3x3 representation in which the unbiased parameters are too inappropriate. The graphs in Figure 52(a) and previously in Figure 49(a) for the Lidstone language model's response to its λ parameter on the TRECVid 2003 and TRECVid 2002 collection show different trends for the DCT 3x3x3x3x3 representation. For the TRECVid 2002 collection its MAP peaks with its parameter $\lambda = 0.02$ and then it declines in performance for larger parameter values, which is quite

different to the other DCT representations on that collection For the TRECVid 2003 collection Lidstone peaks for $\lambda = 0\ 90$ and has a far more consistent response curve compared to the other representations As mentioned previously a single topic image, $topic77_1$, is skewing the results for the DCT 3x3x3x3x3 representation on the TRECVid 2002 collection and thereby invalidating the unbiased results for this representation on the TRECVid 2003 collection Unfortunately, the performance of the DCT 3x3x3x3x3 representation after removing topic 77 is very poor and the resulting optimised parameter value, $\lambda = 8\ 0$, is even more inappropriate for use on the TRECVid 2003 collection Using more similar collections in our experiments, such as TRECVid 2003 with the TRECVid 2004 collection, should reduce this problem as both of these collections are more visually similar

The Jelinek-Mercer language model is again the best optimised interpolation-based language model for all tested representations and peaks, see Table 52 (appendix page 257), at a MAP of 0 0124 for the DCT 8x8x8 texture representation It is also the best unbiased interpolation-based language model for all but the DCT 3x3x3x3x3 representation It achieves the best unbiased MAP of 0 0124 for any retrieval model due to a close correspondence between its optimised smoothing parameter λ for the DCT 8x8x8 representation on both collections The somewhat erratically performing Witten-Bell interpolation smoothing performs next best after Jelinek-Mercer interpolation smoothing for the TRECVid 2003 collection, while Absolute interpolation performs worst

The graphs in Figure 52(d) and previously in Figure 49(g) for the Jelinek-Mercer language model's response to its parameter λ on the TRECVid 2003 and TRECVid 2002 collection show different trends for the DCT 3x3x3x3x3 representation The Jehnek-Mercer language model peaks for very low smoothing, $\lambda = 0\ 01$, on the DCT 3x3x3x3x3 representation relative to other DCT representations for the TRECVid 2002 search task Whereas for the TRECVid 2003 search task the DCT 3x3x3x3x3 representation behaves more consistently with respect to the other texture representations and peaks at $\lambda = 0\ 20$ Unfortunately the unbiased parameter estimate from TRECVid 2002 significantly underestimates the required amount of smoothing when used on the TRECVid 2003 collection In contrast to Lidstone smoothing, removing topic 77 from TRECVid 2002 produces an optimised $\lambda = 0\ 15$ for Jelinek-Mercer smoothing which would produce far better unbiased results on the TRECVid 2003 collection

For the standard visual models, the relative performance order returns to Jensen-Shannon distance followed by Manhattan distance and with Euclidean in last place Jensen-Shannon distance achieves a MAP of 0 0114 for the DCT 8x8x8 representation, see Table 52, which is lower than many of the unbiased language models' results

Overall, the Jelinek-Mercer and Lidstone language models outperform the standard visual retrieval models for global DCT texture Though the DCT 8x8x8 representation performs best for the TRECVid 2003 search task we believe that ultimately the next best representation, DCT 3x3x3x3x3, has better synergetic potential when combined with the other features as it captures more texture or spatial frequency information and therefore it is likely to better complement the colour and edge representations when combined with them The results for DCT 3x3x3x3x3 are shown in Table 53 (appendix page 258) for the TRECVid 2002 and TRECVid 2003 search tasks, which we will now compare with regional variations

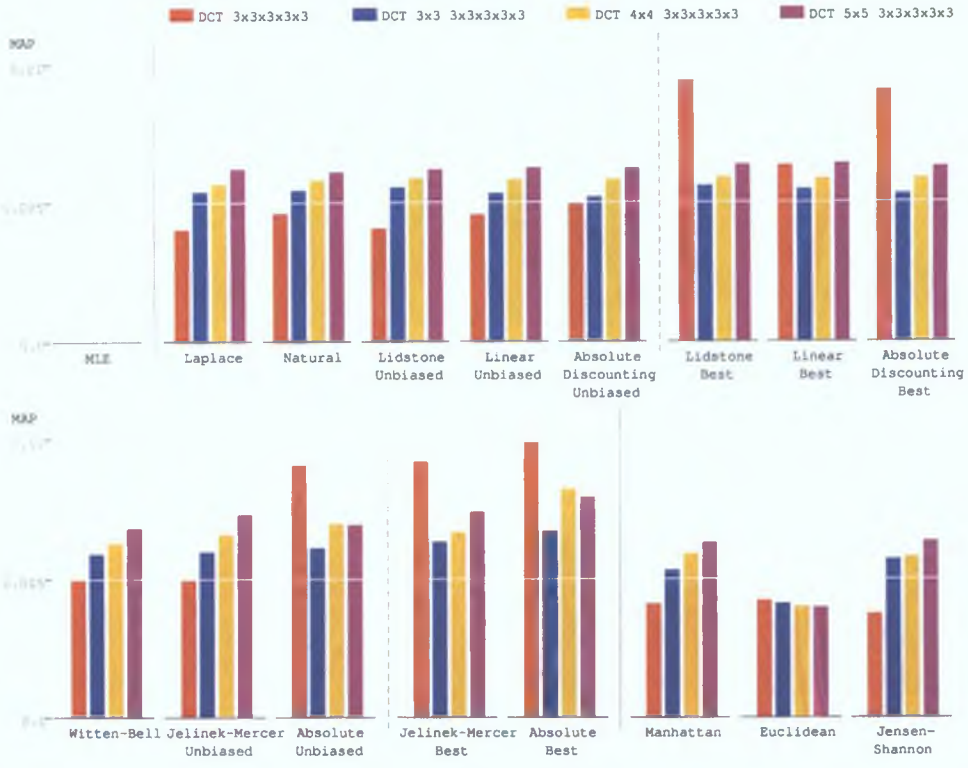
6 5 2 Regional Texture

We will now consider regional texture representations using the DCT 3x3x3x3x3 representation for 3x3, 4x4 and 5x5 regions The results for the TRECVid 2002 and TRECVid 2003 search tasks are shown in Figures 53 and 55 respectively

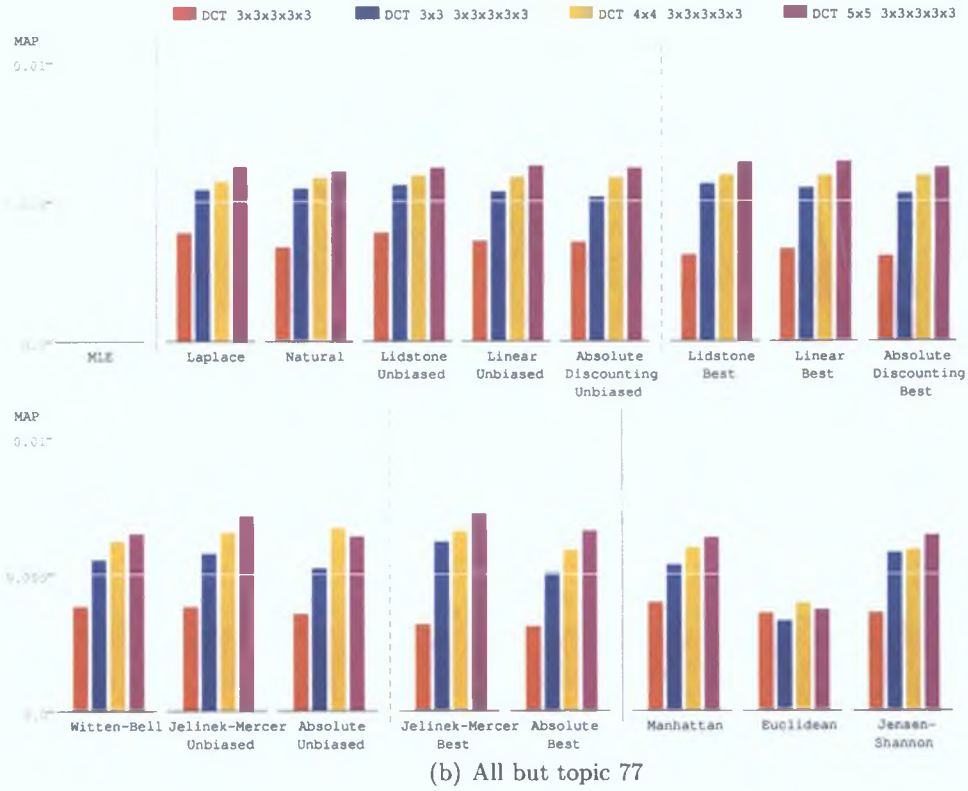
The results for the TRECVid 2002 search task displayed in Figure 53(a) indicate that the regional texture representations perform worse than the non-regional texture representation for the best performing retrieval model, Absolute interpolation, and for the optimised language models As discussed previously, the relatively high MAP for these retrieval models for the non-regional DCT 3x3x3x3x3 texture representation is mainly due to the topic image $topic77_1$

The best 20 topic results for the optimised retrieval model Absolute interpolation, which has the highest MAP for the TRECVid 2002 search task, are shown in Figure 54 The Absolute interpolation language model achieves the highest MAP for the non-regional representation but significantly only two of its top 20 topic results achieve their highest average precision using this non-regional representation The other 18 of the 20 best topic images achieve their highest average precision using the regional representations - 5 for 3x3 regions, 7 for 4x4 regions and 6 for 5x5 regions Unfortunately, as for global texture-based retrieval the average precision for the topic image $topic77_1$ swamps the other topics' average precisions in the calculation of the MAP statistic The results in Table 54 (appendix page 259) compare the 5x5 regional texture representation with the non-regional representation These results show that the MAP for the retrieval models that decrease in performance when using the 5x5 regional representation, Absolute interpolation and the optimised language models, are not statistically significant and actually have very large Wilcoxon p-values greater than 0 8 This table also shows that the improvements in some of the retrieval models for the 5x5 regional texture representation are statistically significant

We present adjusted regional texture results in Figure 53(b) for the TRECVid 2002 search task which disregard topic 77 These adjusted TRECVid 2002 results show that the MAP increases when using regional representations for all retrieval models except for the poorly performing MLE and Euclidean models This outcome for the adjusted results is consistent with our TRECVid 2003 texture results and our previous results for the regional colour and edge features, which also showed that regional representations are more effective The difference in MAP between the regional representations for the adjusted TRECVid 2002 search task is quite small but in general 5x5 regions is better than 4x4 regions, which in turn is better than 3x3 regions The unbiased retrieval models, see Table 54 (appendix page 259), achieve between 26% and 70% better MAP for the 5x5 regional representation than for the non-regional representation But this improvement is not enough to close the gap between the DCT 3x3x3x3x3 representation and the better DCT 4x4x4x4 representation on the TRECVid 2002 collection The 5x5 regional DCT 3x3x3x3x3 texture representation actually underperforms the non-regional DCT 4x4x4x4 representation for the TRECVid 2002 search task We believe, and all evidences suggests, that regional variations of the DCT 4x4x4x4 would improve upon this representation but we have not, as yet, evaluated these regional variations We will now further discuss the results for the 5x5 regional DCT 3x3x3x3x3 representation for the discounted language models, the interpolated language models and the standard visual retrieval models



(a) All topics



(b) All but topic 77

Figure 53: Comparison of *regional DCT representations* (DCT 3x3x3x3x3 for no regions, 3x3, 4x4 and 5x5 regions) using language models and standard visual retrieval models for the *TRECVid 2002* search task over (a) all topics and (b) all but topic 77.

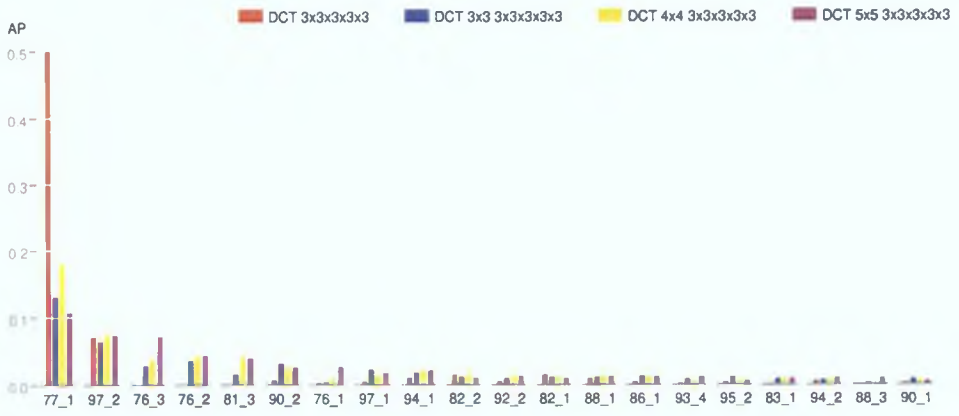


Figure 54: Comparison of *regional DCT representations* (DCT 3x3x3x3x3 for no regions, 3x3, 4x4 and 5x5 regions) *on the 20 most successful TRECVid 2002 topic images* using the optimised Absolute interpolation language model.

As can be seen from Table 54 (appendix page 259) for TRECVid 2002 all the unbiased discounting-based language models achieve the same MAP of 0.0062 for 5x5 regions except for the Natural discounting model which achieves a lower MAP of 0.0061. The optimised discounting-based language models barely improve on this result and the best of which is the optimised Linear model which only achieves a MAP of 0.0064. For the 5x5 regional DCT representation the interpolation-based language models achieve better MAP results than the discounting-based language models and the standard visual retrieval models. Jelinek-Mercer has the best unbiased MAP of 0.0073, while Absolute interpolation has the best optimised MAP of 0.0079. In general, interpolation-based language models seem better than the discounting-based models for regional representations.

The standard visual retrieval models except Euclidean distance achieve similar results to the discounting-based language models with a MAP of 0.0063. The poorly performing Euclidean distance achieves a MAP of 0.0040. Overall the results for regional texture, even though in some cases statistically significantly better than the global representation, are very poor for the TRECVid 2002 search task.

The results for the TRECVid 2003 search task in Figure 55 present a stronger case for the regional texture representations. Many of the retrieval models perform over 100% better on the 5x5 regional representation than on the non-regional representation. All the retrieval models except of course the MLE retrieval model perform better on the regional representation than on the global representation and the best regional representation is 5x5 regions, followed by 4x4 regions, which is in turn followed by 3x3 regions. The only exception is the unbiased Lidstone results where 5x5 regions performs worse than the other regional representations.

The 20 best topic image results for the best retrieval model, the optimised Jelinek-Mercer language model, are shown in Figure 56. It is evident from this chart that no single topic overwhelms the other topic results. The non-regional representation is best for none of the top 20 topic images whereas 5x5 regions is best for 13, 4x4 regions is best for 1 and 3x3 regions is best for the remaining 6 of the top 20 topic images. This is a significant endorsement of the regional texture representation, at least for the top performing query images.

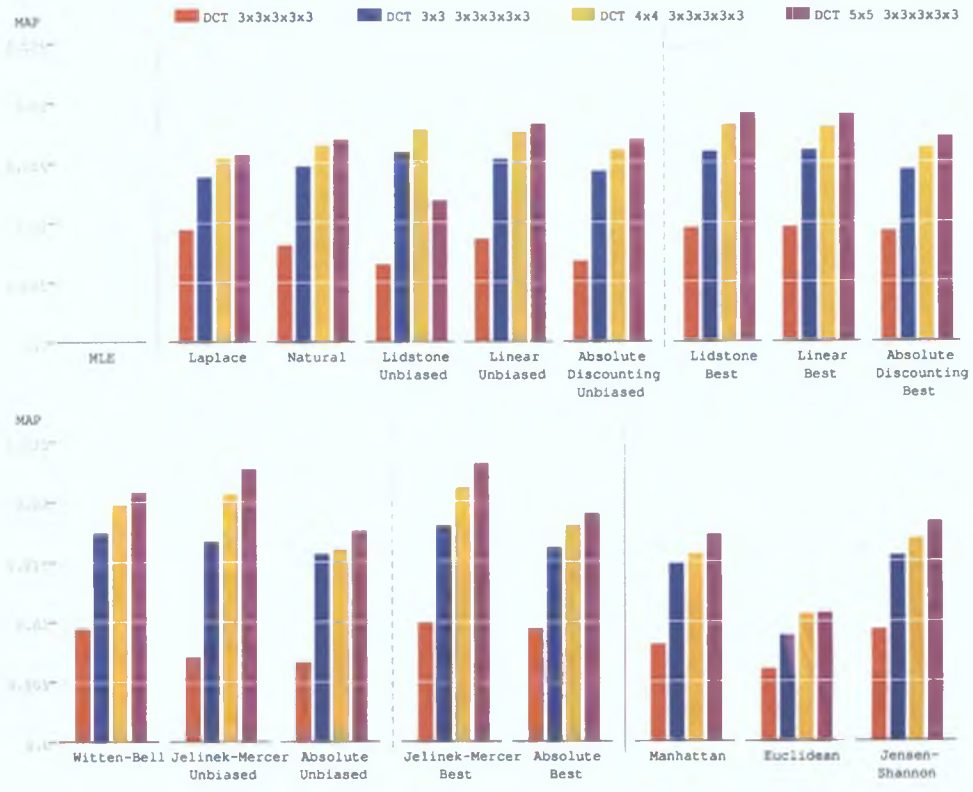


Figure 55: Comparison of *regional DCT representations* (DCT 3x3x3x3x3 for no regions, 3x3, 4x4 and 5x5 regions) using language models and standard visual retrieval models for the *TRECVid 2003* search task.

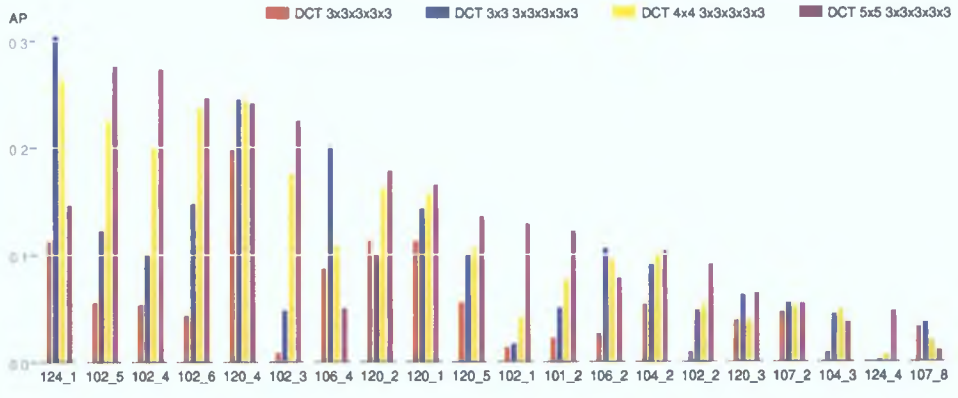


Figure 56: Comparison of *regional DCT representations* (DCT 3x3x3x3x3 for no regions, 3x3, 4x4 and 5x5 regions) *on the 20 most successful TRECVid 2003 topic images* using the optimised Jelinek-Mercer language model.

We compare the retrieval models for the 5x5 regional representation with the other regional and non-regional representations in Table 55 (appendix page 260) The 5x5 regional representation is *statistically significantly* better than the non-regional representation for all retrieval models except of course the defunct MLE language model These positive results are extremely significant with many retrieval models achieving over 100% improvements in MAP and with the related Wilcoxon p-values close to 0 000 after rounding The comparison of the retrieval models for the 5x5 regional texture representation with the 3x3 and 4x4 regional representations also indicate many significant improvements except for the unbiased Lidstone language model The results for the 5x5 regional representation are in general better than for the 3x3 regional representation by between 12% and 35% and are also between 2% and 10% better than for the 4x4 regional representation Even many of the improvements between the 5x5 regional representation and the 4x4 regional representation are statistically significant

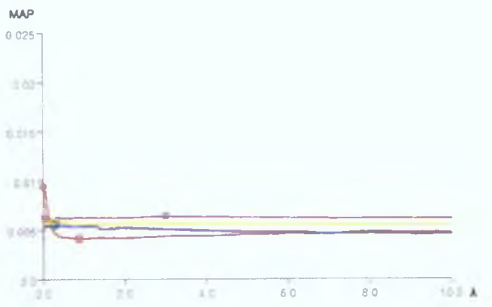
The response of the parametric language models' MAP to different smoothing parameter values are shown in Figure 57 The graphs for TRECVid 2003 show that there is a strict performance order for the regional texture representations for nearly all the parameter values of the different parametric language models - 5x5 regions is best and is followed by 4x4 regions which is followed by 3x3 regions and then by the global texture representation This pattern can also be observed for the TRECVid 2002 parametric language models, also shown in Figure 57, but due to the influence of the topic image $topic77_1$ the global texture representation achieves higher MAP than the regional representations for low smoothing parameter values We will now discuss in more detail the results for the language models and the visual retrieval models using the 5x5 regional texture representation

The Lidstone and Linear language models are the best performing discounting-based language models on the TRECVid 2003 collection with an optimised MAP of 0 0191 and 0 0190 respectively for the 5x5 regional texture representation The Linear model has a far better unbiased result than the Lidstone model but it is hard to take much cognisance of these cross validated results due to the poor performance of the retrieval models for the DCT 5x5 3x3x3x3x3 regional texture representation on the TRECVid 2002 collection

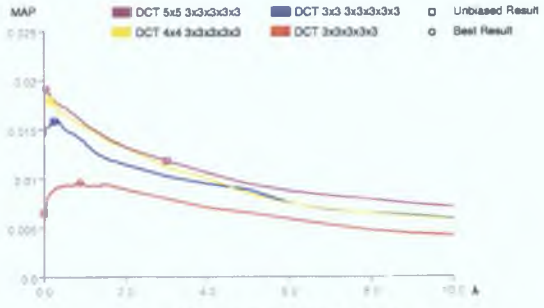
The Jelinek-Mercer language model, which achieves an optimised MAP of 0 0232 and an unbiased MAP of 0 0226, is clearly the best retrieval model for the regional texture representation on the TRECVid 2003 collection The Absolute interpolation language model achieves much poorer results, which are at the same level as the best discounting-based language models Witten-Bell interpolation achieves results that are mid-way between Absolute interpolation and Jelinek-Mercer interpolation Like most regional feature representations, the regional texture representation seems to be better modelled for the search task by using a simple interpolation with the background collection model (Jelinek-Mercer smoothing) than by using an interpolation with a uniform noise source (Lidstone smoothing)

Jensen-Shannon distance is the best performing standard visual retrieval model but achieves only a MAP of 0 0183, which is 20% lower than for Jelmek-Mercer smoothing Manhattan distance is the next best and as usual Euclidean distance is the worst standard visual retrieval model

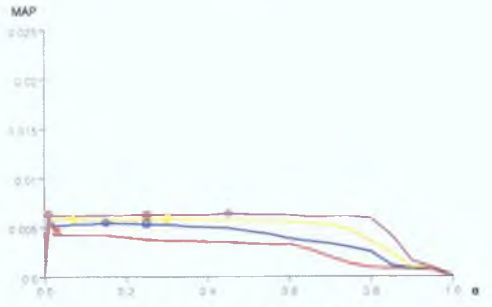
We present statistical significance comparisons of the unbiased retrieval models (except the



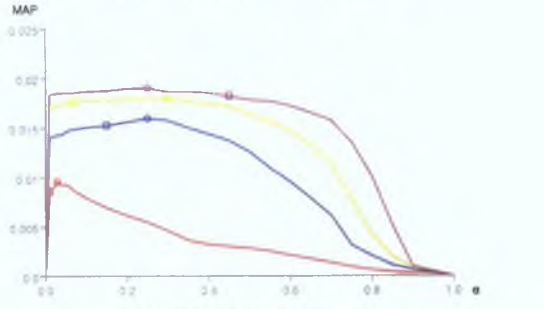
(a) TRECVid 2002: Lidstone



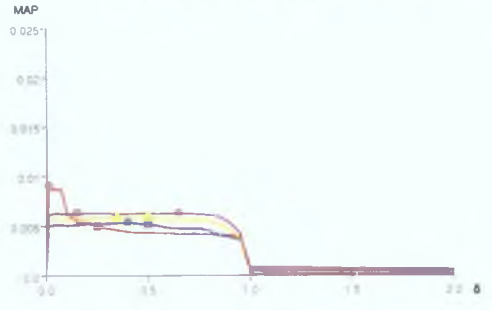
(b) TRECVid 2003: Lidstone



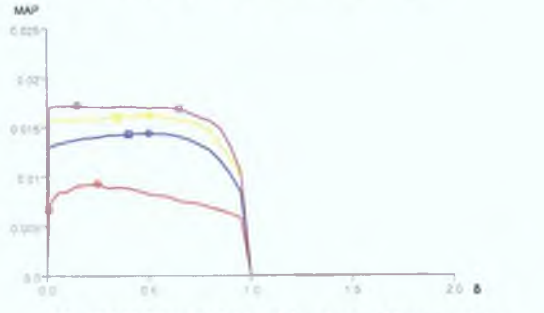
(c) TRECVid 2002: Linear



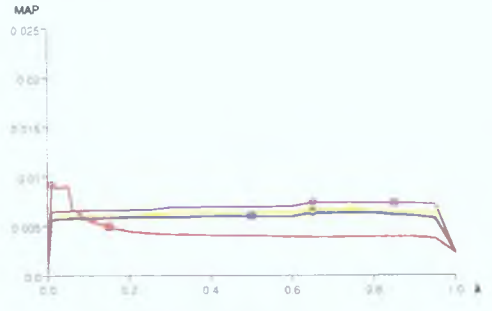
(d) TRECVid 2003: Linear



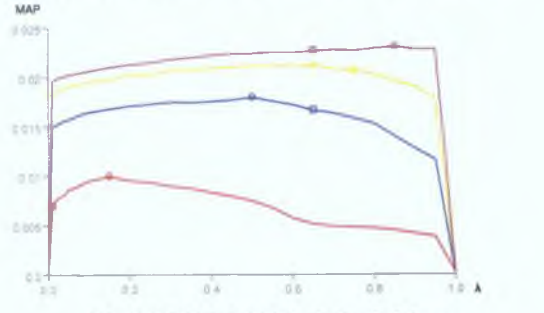
(e) TRECVid 2002: Absolute Discounting



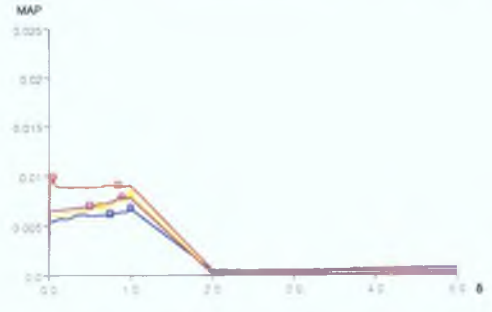
(f) TRECVid 2003: Absolute Discounting



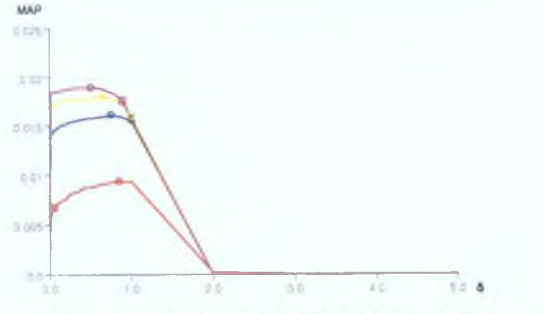
(g) TRECVid 2002: Jelinek-Mercer



(h) TRECVid 2003: Jelinek-Mercer



(i) TRECVid 2002: Absolute Interpolation



(j) TRECVid 2003: Absolute Interpolation

Figure 57: Comparison of MAP for the different parameter values of the parametric language models (Lidstone, Linear, Absolute discounting, Jelinek-Mercer and Absolute interpolation) using the *regional DCT representations* (DCT 3x3x3x3x3 for no regions, 3x3, 4x4 and 5x5 regions) for the *TRECVid 2002* and *TRECVid 2003* search tasks.

Table 17: Statistical significance comparison of retrieval models on the regional 5x5 DCT 3x3x3x3x3 texture feature for the aggregated TRECVid 2002, 2003 and 2004 search tasks.

Ret. Meth.	*MAP/%Dif*	*Wilcoxon Test Results*								
Witten-Bell (WB)		-	>JM	>Abs	>Lin	>JS	>AbsD	>Man	>Lap	>Lid
(WB)	0.0130		.294	.003	.020	.032	.000	.024	.000	.000
Jelinek-Mercer (JM)		<WB	-	>Abs	>Lin	>JS	>AbsD	>Man	>Lap	>Lid
(JM)	0.0127/-2.4%	.294		.272	.246	.295	.157	.256	.000	.013
Absolute (Abs)		<WB	<JM	-	>Lin	>JS	>AbsD	>Man	>Lap	>Lid
(Abs)	0.0118/-9.2%	.003	.272		.418	.382	.051	.246	.023	.006
Linear (Lin)		<WB	<JM	<Abs	-	>JS	>AbsD	>Man	>Lap	>Lid
(Lin)	0.0118/-9.5%	.020	.246	.418		.788	.139	.748	.000	.000
Jensen-Shannon (JS)		<WB	<JM	<Abs	<Lin	-	>AbsD	>Man	>Lap	>Lid
(JS)	0.0116/-10.7%	.032	.295	.382	.788		.038	.386	.000	.000
Absolute Disc. (AbsD)		<WB	<JM	<Abs	<Lin	<JS	-	>Man	>Lap	>Lid
(AbsD)	0.0114/-12.2%	.000	.157	.051	.139	.038		.956	.084	.020
Manhattan (Man)		<WB	<JM	<Abs	<Lin	<JS	<AbsD	-	>Lap	>Lid
(Man)	0.0114/-12.3%	.024	.256	.246	.748	.386	.956		.000	.000
Laplace (Lap)		<WB	<JM	<Abs	<Lin	<JS	<AbsD	<Man	-	>Lid
(Lap)	0.0094/-28.0%	.000	.000	.023	.000	.000	.084	.000		.253
Lidstone (Lid)		<WB	<JM	<Abs	<Lin	<JS	<AbsD	<Man	<Lap	-
(Lid)	0.0093/-28.7%	.000	.013	.006	.000	.000	.020	.000	.253	

poorly performing MLE and Euclidean distance models) on the *DCT 5x5 3x3x3x3x3* feature for the set of topic results across the three TRECVid collections in Table 17. This table indicates that the Witten-Bell language model is overall the best retrieval model for this feature. The Jelinek-Mercer language model again performs very strongly and is statistically insignificantly lower than Witten-Bell by only 2.4%. Similar to the regional edge feature, Lidstone smoothing is worst (excluding the poor MLE and Euclidean from our comparison). The top three retrieval models are interpolation-based language models, which implies that smoothing using the background collection is important for this feature. Jensen-Shannon distance is fifth best according to overall MAP and is surprisingly statistically significantly lower than the Witten-Bell smoothed language model. The p-value is slightly high at 0.032 and therefore we may take this statistical significance relationship with a bit of caution. Because the DCT feature is calculated over 8x8 blocks (not for individual pixels) this feature is rather sparser than previous features, which were calculated for each pixel in the image. It is possible that smoothing using the background distribution is more beneficial than combining document and query as is done for Jensen-Shannon distance when the number of samples is far smaller than the size of the visual language. We need to experiment with far larger visual languages in order to see if there is such a relationship.

We tabulate the results of our tested retrieval models on the official TRECVid multi-example topics in Table 18 that fused the single image results for the *DCT 5x5 3x3x3x3x3* feature using CombSUM. After fusion the Jelinek-Mercer smoothed language model is the overall best retrieval model in terms of MAP. This feature is slightly better than the edge feature for TRECVid 2002, slightly worse for TRECVid 2004, and double the effectiveness of the regional edge feature for TRECVid 2003. In seeking to make this feature cover more DCT bands we have decreased its effectiveness in the hope that it will be useful when fused with colour. The choice of feature representation may have been a poor choice and looks likely overly biased towards the TRECVid

Table 18: Comparison of retrieval models on the DCT 5x5 3x3x3x3x3 feature for official TRECVid topics (i.e. fused topic examples).

VisExsTexture-CombScore	*TRECVid 2002*				*TRECVid 2003*				*TRECVid 2004*				*TRECVid 02-04*			
Retrieval Method	*MAP*	*P10*	*P30*	*P100*	*MAP*	*P10*	*P30*	*P100*	*MAP*	*P10*	*P30*	*P100*	*MAP*	*P10*	*P30*	*P100*
Laplace	**.0095**	.028	.028	.016	**.0312**	.080	.065	.044	**.0030**	.004	.007	.006	**.0149**	.038	.034	.022
Best	**(.0096**	.028	.028	.016)	**(.0331**	.080	.065	.044)	**(.0030**	.004	.007	.006)	**(.0155**	.038	.034	.022)
Lidstone	**.0099**	.044	.029	.018	**.0150**	.028	.033	.028	**.0095**	.039	.019	.019	**.0115**	.037	.027	.022
Best	**(.0099**	.032	.020	.015)	**(.0357**	.108	.083	.057)	**(.0095**	.039	.019	.019)	**(.0186**	.060	.041	.031)
Linear	**.0100**	.048	.028	.020	**.0341**	.100	.076	.052	**.0099**	.035	.020	.015	**.0182**	.062	.042	.029
Best	**(.0101**	.036	.029	.019)	**(.0356**	.104	.084	.054)	**(.0101**	.039	.019	.019)	**(.0188**	.060	.045	.031)
Absolute Discounting	**.0086**	.048	.032	.020	**.0295**	.088	.072	.051	**.0106**	.030	.028	.017	**.0164**	.056	.044	.030
Best	**(.0102**	.048	.032	.021)	**(.0304**	.104	.079	.052)	**(.0107**	.030	.028	.019)	**(.0173**	.062	.047	.031)
Witten-Bell	**.0112**	.052	.033	.021	**.0370**	.116	.088	.060	**.0111**	.035	.025	.020	**.0200**	.068	.049	.034
Best	**(.0120**	.056	.032	.022)	**(.0378**	.108	.093	.059)	**(.0121**	.030	.025	.017)	**(.0209**	.066	.051	.033)
Jelinek-Mercer	**.0142**	.032	.028	.020	**.0417**	.116	.088	.061	**.0057**	.030	.014	.009	**.0209**	.060	.044	.031
Best	**(.0142**	.036	.027	.020)	**(.0422**	.108	.083	.061)	**(.0062**	.030	.014	.008)	**(.0213**	.059	.042	.031)
Absolute	**.0115**	.056	.033	.022	**.0309**	.100	.092	.055	**.0114**	.026	.026	.023	**.0181**	.062	.051	.034
Best	**(.0130**	.048	.031	.020)	**(.0339**	.088	.085	.056)	**(.0119**	.026	.028	.018)	**(.0198**	.055	.048	.032)
Manhattan	**.0114**	.044	.024	.018	**.0312**	.084	.073	.050	**.0117**	.035	.025	.015	**.0183**	.055	.041	.028
Best	**(.0121**	.044	.027	.017)	**(.0329**	.092	.076	.052)	**(.0117**	.035	.025	.015)	**(.0191**	.058	.043	.028)
Jensen Shannon	**.0112**	.036	.025	.018	**.0340**	.104	.077	.054	**.0092**	.026	.019	.012	**.0184**	.056	.041	.028
Best	**(.0114**	.036	.025	.018)	**(.0344**	.096	.079	.053)	**(.0100**	.030	.022	.020)	**(.0189**	.055	.042	.030)
Euclidean	**.0049**	.032	.017	.016	**.0196**	.056	.052	.037	**.0080**	.026	.019	.014	**.0109**	.038	.030	.023
Best	**(.0053**	.032	.019	.018)	**(.0198**	.052	.053	.036)	**(.0089**	.030	.022	.015)	**(.0114**	.038	.032	.023)
Average of unbiased	**.0103**	.042	.028	.019	**.0304**	.087	.072	.049	**.0090**	.029	.020	.015	**.0168**	.053	.040	.028
Average of best	**.0108**	.040	.027	.018	**.0336**	.094	.078	.052	**.0094**	.030	.021	.015	**.0182**	.055	.042	.029

2003 collection In comparison with the DCT GMM model, the Jelinek-Mercer language model achieves 50 5% lower MAP on TRECVid 2002, 34 5% better MAP on TRECVid 2003 and 43 0% lower MAP on TRECVid 2004 The DCT GMM model contains colour information as well as texture information, which may account for most of its performance We have separated colour and texture into different features and we have previously shown that the regional colour feature alone is better than the DCT GMM model on the TRECVid 2003 and TRECVid 2004 search tasks

6.6 Summary

As has been demonstrated in this chapter, the language modelling approach to text information retrieval can be successfully applied to the visual retrieval task for a range of different visual features Our test cases involved colour, edge and texture based features The choice of discrete representation for each feature needs to be chosen with care and by experimentation

The MPEG7 inspired HSV 16x4x4 colour representation was the best of the tested HSV colour representations The Canny 64+1 was the best of the tested Canny representations, though Canny 32+1 produced quite similar results and a higher number of quantisation levels may produce slightly better results but we did not test this There was disagreement on which DCT texture representation was best on both test collections For TRECVid 2002 the best representation was DCT 4x4x4x4 whereas for TRECVid 2003 the best representation was DCT 8x8x8 In the end we chose DCT 3x3x3x3x3 as a compromise that took into account the experimental results and sought to maximise the amount of frequency information that would be captured in the texture representation

The 5x5 regional representations seem to have the best number of regions out of the tested regional representations More regions may be more effective for TRECVid 2003 and TRECVid 2004 search tasks but this would require more storage and a slower retrieval time

In terms of traditional visual retrieval models Jensen-Shannon distance seems to be the best with Manhattan distance in second place Euclidean distance seems to consistently perform worse than these methods The use of Euclidean distance in (TRECVid) video retrieval seems ill founded The less used Jensen-Shannon distance seems to be a strong retrieval model for the video shot retrieval task, at least for the tested features

As expected the MLE query-likelihood model performs very badly without smoothing The structural smoothing that combines the shot representation with adjacent shots, stories or the video units was not beneficial to video retrieval for the TRECVid 2003 television news content or for the majority of topics on the TRECVid 2002 collection

The results for visual features using global representations did not show much consistent difference between the discounting and the interpolation-based language models However, for 5x5 regions we found that Jelinek-Mercer smoothing in general outperforms the discounting methods The regional HSV colour feature is the exception, and perhaps because of its more uniform distribution Lidstone smoothing is superior, though the difference is not statistically significant

For the parametric language models the choice of smoothing parameter value can have a divisive effect on retrieval performance Our experiments on just two types of collections (television news for TRECVid 2003 and TRECVid 2004 and the old poorly encoded miscellaneous public domain videos for TRECVid 2002) limits the research we can do on this topic As TRECVid 2003 and TRECVid 2004 video collections are so similar it would be interesting to cross-validate the parametric language models on both collections, which may result in more consistent and clearer results for TRECVid 2003

Another source of concern arising from our results is the problem in using MAP for optimising results and we should try to identify a better optimisation criteria There is much confusion and extra investigation necessary when evaluating visual-based retrieval results due to their overall poor performance and the ability of one or two topic images to dominate the MAP and other standard IR measures thereby complicating the interpretation, comparison and training of retrieval models

Overall we found that the discrete visual language modelling approach can be applied successfully to the visual video shot retrieval task and it is competitive with the best of the standard visual retrieval models

CHAPTER VII

EVALUATION III VIDEO RETRIEVAL USING COMBINED FEATURES

In this chapter, we investigate fusion methods for combining the discrete language models for the text, colour, edge and texture features for the video shot retrieval task We evaluate fusion methods separately for combining multiple visual features, multiple visual examples and the multiple modalities using fully automatic experiments on the TRECVid search tasks We compare fusion methods based on normalised score and normalised rank that use either the average, weighted average or maximum of individual results as their combination function We also compare these results with a simple probabilistic combination that assumes all features and examples are fully independent We find that combinations based on scores are as good as or better than combinations based on rank and independent probability, and that the weighted average function is best for combining visual features and modalities, while the maximum score function is best for combining multiple visual examples

7 1 Introduction

7 2 Experiments with multiple visual features

7 3 Experiments with multiple visual examples

7 4 Experiments with multiple modalities

7 5 Discussion

7 6 Summary

7.1 Introduction

In this chapter we experiment with the fusion of retrieval results in order to combine (A) multiple visual features, (B) multiple visual examples and (C) the multiple modalities text and visual

We perform fully automatic fusion experiments on the TRECVid 2002, 2003 and 2004 collections using the previously established optimised and unbiased retrieval results for the text and visual features In contrast to our TRECVid 2002 and 2003 results, the optimised fusion experiments in this chapter for the TRECVid 2004 collection represent the optimised combination of the *unbiased* visual features' results since, as yet, we have no optimised results for the visual features on this collection

We have chosen to investigate the fusion of only one type of text retrieval model, the hierarchical Jelinek-Mercer *shot+adj+video* language model, with the visual results due to its superior performance over the other tested language models and also in order to simplify the presentation and discussion of our results We did not choose the better semantic *shot+story* hierarchical structure as this is was only available to us for the TRECVid 2003 collection The key problem

in this chapter is not which text retrieval model to choose, but how to combine the results for visual features and visual examples successfully with each other and with the text results

We experiment with combining features from a wide selection of visual retrieval models, such as discounting-based language models, interpolation-based language models, and standard visual retrieval models, in order to achieve a wider and hopefully more balanced view of the benefits and faults of the different fusion methods For each of the combination tasks we present an overview of the fusion method's average performance across all these visual retrieval models These overview averages, while not attributable to any one fusion result, provide an indication of the overall performance of the fusion methods on each TRECVid collection

We concentrate our fusion experiments on combining the visual results for the 5x5 regional feature representations for all visual retrieval models except the generally poor performing MLE, Natural and Euclidean models It would be interesting in future work to repeat all our fusion experiments on the global and other regional feature representations in order to get a better indication of relative benefits of the 5x5 regional representation

The rest of this chapter is organised as follows Section 2 deals with experiments for combining multiple visual features, Section 3 presents experiments for combining multiple visual examples, Section 4 presents experiments for combining the different modalities text and visual and Section 5 discusses the fusion results in particular from the point of view of the underlying visual retrieval models We will end this chapter with a short summary of the results

7.2 Experiments with multiple visual features

In this section we evaluate the fusion methods *Vis-CombJointPr* (sum log-likelihoods), *Vis-CombWtRank* (weighted Borda count) and *Vis-CombWtScore* (weighted interpolation of normalised scores), which we described in Chapter 4, for combining the visual results of the colour, edge and texture features

We summarise the fusion results across all retrieval models for combining the visual features in Table 19 for the TRECVid 2002, 2003 and 2004 search tasks The fusion results for TRECVid 2002, see Table 19(a), are very poor with all but the *Vis-CombJointPr* fusion method producing worse unbiased results than for the colour feature alone The *Vis-CombJointPr* fusion produces the best unbiased fusion, which is 3 7% better than the colour-only results The results for TRECVid 2003 in Table 19(b) indicate poor performance for the *Vis-CombWtRank* fusion methods but the *Vis-CombWtScore* fusion method is the best and achieves on average a 3 2% improvement over the colour-only results for the unbiased results The unbiased results are not encouraging for combining multiple visual features for both the TRECVid 2002 and 2003 collections, which may be due to the fact that they are tuned on each others results but both collections are very different In contrast to the other TRECVid collections, the results for TRECVid 2004 in Table 19(c) show positive improvement for all fusion methods in comparison to the individual feature results This is probably due to the more even performance of the colour-only and texture-only retrieval models on the TRECVid 2004 collection than on the other two collections The *Vis-CombWtScore* fusion method is again on average the best unbiased fusion method with an improvement of 28 5% over the colour-only results, while the *Vis-CombJointPr*

Table 19: Comparison of the average results across retrieval models for combining the colour, edge and texture results using the *Vis-CombJointPr*, *Vis-CombWtRank* and *Vis-CombWtScore* fusion methods on the (a) TRECVid 2002, (b) TRECVid 2003 and (c) TRECVid 2004 collections. The colour, edge and texture results are also shown here for comparison and the percentage improvement column is relative to the colour results.

(a) TRECVid 2002 average results across retrieval models.

TRECVid 2002	*MAP*	*P@10*	*P@30*	*P@100*	*Impr.*
Colour	**.0158**	.041	.027	.018	
Optimised	(**.0159**	.043	.027	.017)	
Edge	**.0066**	.033	.021	.014	
Optimised	(**.0073**	.033	.021	.015)	
Textures	**.0065**	.026	.018	.015	
Optimised	(**.0069**	.025	.019	.015)	
Vis-CombJointPr	**.0158**	.042	.027	.018	+3.7%
Optimised	**.0162**	.043	.027	.017	+2.9%
Vis-CombWtRank	**.0137**	.044	.030	.019	-13.5%
Optimised	(**.0170**	.045	.028	.018)	+4.6%
Vis-CombWtScore	**.0155**	.054	.033	.020	-2.6%
Optimised	(**.0172**	.047	.030	.019)	+6.0%

(b) TRECVid 2003 average results across retrieval models.

TRECVid 2003	*MAP*	*P@10*	*P@30*	*P@100*	*Impr.*
Colour	**.0259**	.089	.069	.040	
Optimised	(**.0269**	.092	.071	.041)	
Edge	**.0100**	.043	.034	.022	
Optimised	(**.0108**	.044	.036	.023)	
Textures	**.0176**	.066	.053	.034	
Optimised	(**.0194**	.071	.059	.037)	
Vis-CombJointPr	**.0265**	.088	.070	.042	+2.1%
Optimised	**.0272**	.091	.071	.042	+1.2%
Vis-CombWtRank	**.0255**	.089	.071	.041	-1.4%
Optimised	(**.0272**	.090	.072	.043)	+2.3%
Vis-CombWtScore	**.0267**	.089	.072	.043	+3.2%
Optimised	(**.0290**	.093	.076	.047)	+9.0%

(c) TRECVid 2004 average results across retrieval models.

TRECVid 2004	*MAP*	*P@10*	*P@30*	*P@100*	*Impr.*
Colour	**.0087**	.036	.025	.014	
Edge	**.0050**	.018	.014	.009	
Textures	**.0080**	.027	.018	.013	
Vis-CombJointPr	**.0092**	.037	.026	.015	+4.2%
Vis-CombWtRank	**.0100**	.039	.026	.015	+14.7%
Optimised	(**.0131**	.041	.030	.017)	+50.2%
Vis-CombWtScore	**.0112**	.041	.028	.016	+28.5%
Optimised	(**.0127**	.042	.029	.017)	+45.8%

and the *Vis-CombWtRank* fusion methods achieve an improvement of on average +4 2% and +14 7% respectively

Table 19 also summarises the results for the optimised fusion methods The *Vis-CombWtScore* fusion method on average achieves the best optimised results for TRECVid 2002 and TRECVid 2003 of +6 0% and 9 0% respectively, while the *Vis-CombWtRank* achieves on average the best optimised results for TRECVid 2004 with an improvement of +50 2% compared to *Vis-CombWtScore*'s slightly smaller improvement of +48 8% over the colour-only result

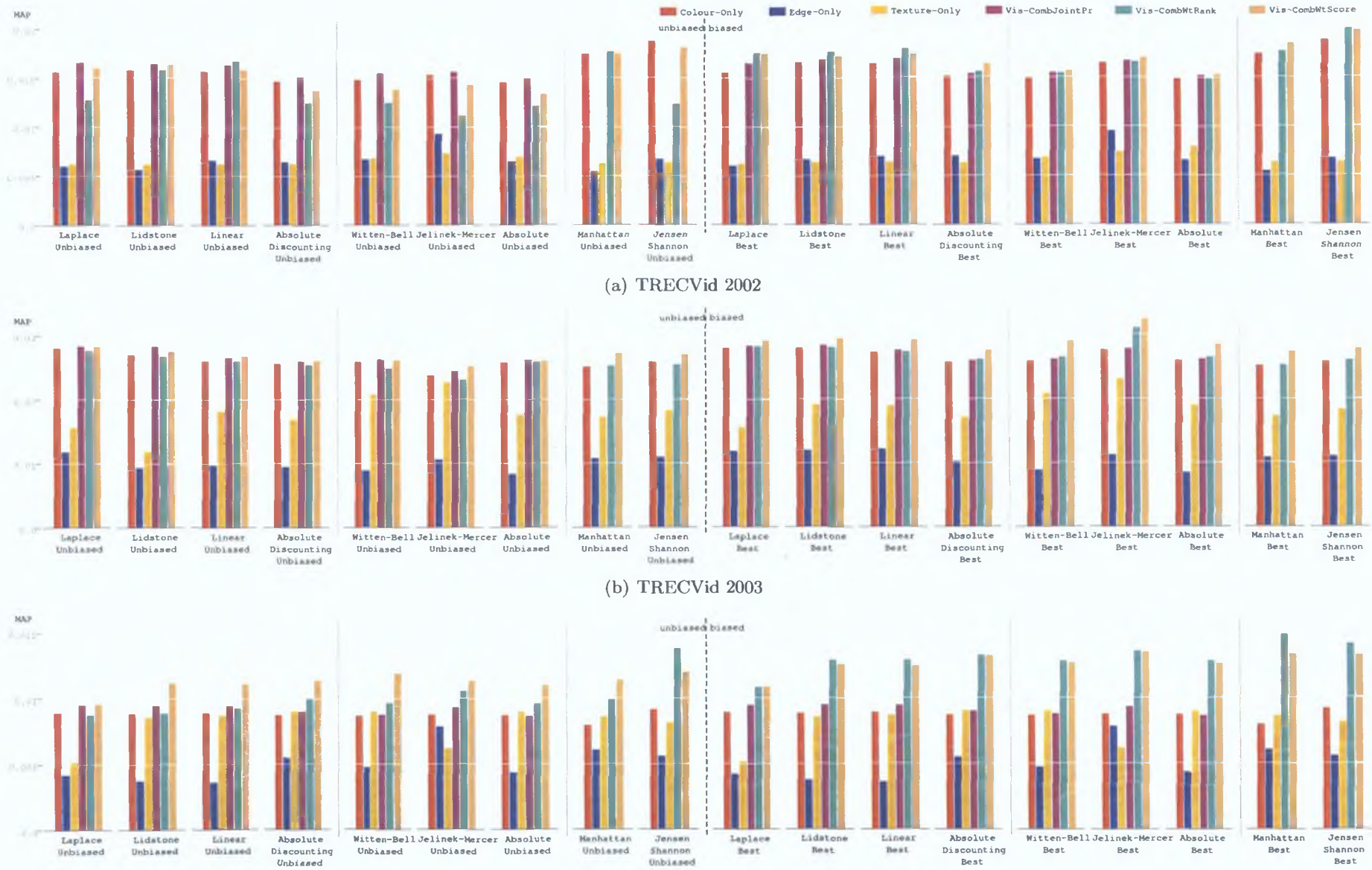
We conclude from these results for the three TRECVid collections that the *Vis-CombWtScore* fusion method is better on average for combining the colour, edge and texture retrieval results than the other fusion methods We will now examine the results of these fusion methods for the different retrieval models as displayed in Figure 58 This figure also displays the individual performance of the colour, edge and texture visual features that are inputs into the fusion methods

We will now discuss in more detail the TRECVid 2002 results as shown in Figure 58(a), which show disappointing unbiased results for combining the individual visual features using the fusion methods *Vis-CombWtRank* and *Vis-CombWtScore* for all visual retrieval models The *Vis-CombJointPr* fusion method is slightly better than the colour results alone for all retrieval models The *Vis-CombWtScore* fusion method is in general the next best fusion method for most of the unbiased results The optimised results are less clear cut between *Vis-CombWtScore* and *Vis-CombWtRank* and show that both methods achieve similar potential performance with one or the other slightly better on each occasion for the different retrieval results

We compare the results of the *Vis-CombWtScore* fusion method with the results for colour-only, *Vis-CombJointPr* and *Vis-CombWtRank* fusion methods on the TRECVid 2002 search task in Table 59 (appendix page 265) We can see from this table that the *Vis-CombWtScore* fusion results are statistically significantly better than many of the *Vis-CombWtRank* fusion results for the different visual retrieval models The *Vis-CombWtScore* fusion method is also statistically significantly better than the colour-only results for nearly all the optimised results The small improvements for the *Vis-CombJointPr* over the *Vis-CombWtScore* fusion method are never statistically significant for any of the unbiased results

The optimised parameters for the *Vis-CombWtScore* fusion method on the TRECVid 2002 collection in Table 59 (appendix page 265) show that the inputs to this fusion method, colour, edge and texture, have non-zero weights associated with them for most of the visual retrieval models The exceptions are the combined results of Absolute discounting, Witten-Bell and Absolute interpolation where the weight for texture is zero, which indicates a failure of the *Vis-CombWtScore fusion method to combine the* texture results with the more successful colour results This problem of texture not combining with the other features on the TRECVid 2002 collection is more prevalent for the *Vis-CombWtRank* fusion method where for most retrieval models the optimised weight of texture is zero

We will now discuss in more detail the TRECVid 2003 results as shown in Figure 58(b) We can see from this figure that the three fusion methods achieve quite similar results for fusing the different visual retrieval results *Vis-CombWtScore* more often than not is better than



(a) TRECVid 2002

(b) TRECVid 2003

(c) TRECVid 2004

Figure 58: Comparison of the fusion methods *Vis-CombJointPr*, *Vis-CombWtRank* and *Vis-CombWtScore* for combining the colour, edge and texture results for the different retrieval models on the (a) TRECVid 2002, (b) TRECVid 2003 and (c) TRECVid 2004 collections. The *colour-only*, *edge-only* and *texture-only* results are also displayed for comparison.

Vis-CombJointPr, while *Vis-CombWtRank* is consistently the worst for the unbiased fusion results We compare the *Vis-CombWtScore* fusion method with the colour-only and other fusion methods in Table 60 (appendix page 266) All the *Vis-CombWtScore* fusion results are statistically significantly better than the colour-only results except for the unbiased combination of the Witten-Bell and Absolute interpolation's results, which also improve on the colour-only results but are not statistically significant The comparison of the *Vis-CombWtScore* fusion method with the *Vis-CombWtRank* fusion method is equally positive with all results better using the *Vis-CombWtScore* fusion method and most of these improvements are also statistically significant Interestingly, when *Vis-CombWtScore* is better than *Vis-CombJointPr* the difference is statistically significant, but when *Vis-CombWtScore* is worse than *Vis-CombJointPr* the difference is not statistically significant Unlike TRECVid 2002, the optimised *Vis-CombWtScore* and *Vis-CombWtRank* fusion methods on the TRECVid 2003 collection never exclude any feature with every feature getting a non-zero weight

We will now consider in more detail the TRECVid 2004 results as shown in Figure 58(c) The unbiased results for TRECVid 2004 show that the *Vis-CombWtScore* is better than the other two fusion methods for combining all visual feature's results from the different retrieval models All fusion methods improve on the colour-only results by a larger magnitude than was seen on the other two collections The relative strength of the fusion methods on the TRECVid 2004 collection is related to the more even performance of the visual features on this collection – on the TRECVid 2002 and TRECVid 2003 collections the texture-only retrieval models on average achieve respectively 41% and 68% of the MAP of the colour-only retrieval models, while on the TRECVid 2004 collection the texture-only results are on average 92% of the colour-only result The edge-only results are also improved on the TRECVid 2004 collection (57% of colour) compared to the TRECVid 2002 (41% of colour) and TRECVid 2003 (38% of colour) results

The results for the *Vis-CombWtScore* fusion method in Table 61 (appendix page 267) for TRECVid 2004 show that it is on average 28 5% better than the colour-only results, only +21 0% better than the *Vis-CombJointPr* fusion results and +13 1% better than the *Vis-CombWtRank* fusion results The *Vis-CombWtScore* fusion results for all retrieval models achieve improvements in their MAP compared to the colour-only results The improvements range from 7 4% for the poorly performing Laplace language model to 43 6% for combining the Manhattan distance's results The fusion of Jensen-Shannon distances's colour, edge and texture results achieve the highest MAP of 0 0120 (+31 3%) The *Vis-CombWtScore* fusion method is better than the *Vis-CombWtRank* fusion method for all unbiased combinations of retrieval models except for the combination of the Jensen-Shannon distance's results The improvements of *Vis-CombWtScore* compared to *Vis-CombJointPr* are never statistically significant, which is likely due to the features having more even performance on this collection

We will now discuss the performance of the *Vis-CombWtScore* fusion method in terms of the underlying visual retrieval models for the three TRECVid collections We synthesise the results for these collections into a single ordering of retrieval models based on their average percentage decrease in MAP compared to the best retrieval result on each collection in Table 20 (see specifically the "Average %Dif" column), which is based on Tables 62, 63 and 64 (appendix pp 268, 268 and 269) In this table we also display other ad hoc measures of performance such as the average rank of the retrieval models and the average number of statistically significantly better retrieval models on the three TRECVid collections (see specifically the "R(S)" column)

Table 20 Comparison of unbiased retrieval models for the TRECVid 2002, 2003 and 2004 search tasks using the *Vis-CombWtScore* fusion method for combining colour, edge and texture results using three performance measures %Dif = difference from best result's MAP (lower better) R = rank of retrieval models based on MAP (lower better) S = number of statistically significantly better retrieval models (lower better)

	TRECVid'02		*TRECVid'03*		*TRECVid'04*		*Average*	
Ret Method	*%Dif*	*R(S)*	*%Dif*	*R(S)*	*%Dif*	*R(S)*	*%Dif*	*R(S)*
Jensen-Shannon	0%	1(0)	-3 9%	4(0)	0 0%	1(0)	**-1 3%**	**2 0(0 0)**
Manhattan	-3 3%	2(1)	-3 2%	3(0)	-5 0%	3(1)	**-3 8%**	**2 7(0 7)**
Lidstone	-9 4%	3(0)	-1 5%	2(0)	-7 5%	6(0)	**-6 1%**	**3 7(0 0)**
Linear	-12 7%	5(1)	-5 0%	5(2)	-7 5%	7(0)	**-8 4%**	**5 7(1 0)**
Laplace	-11 6%	4(1)	0%	1(0)	-20 0%	9(3)	**-10 5%**	**4 7(1 3)**
Witten-Bell	-23 2%	7(0)	-7 0%	6(5)	-0 8%	2(0)	**-10 3%**	**5 0(2 6)**
Absolute Discounting	-24 3%	8(6)	-7 8%	8(5)	-5 0%	4(1)	**-12 4%**	**6 7(4 0)**
Jelinek-Mercer	-21 0%	6(0)	-10 6%	9(2)	-5 8%	5(0)	**-12 5%**	**6 7(0 7)**
Absolute	-26 0%	9(4)	-7 4%	7(6)	-8 3%	8(0)	**-13 9%**	**8 0(3 3)**

We can see from these synthesised results that the *Vis-CombWtScore* fusion of the Jensen-Shannon distance's colour, edge and texture results is the best overall fusion result in terms of our three ad hoc performance measures – Jensen-Shannon distance's results are on average ranked second for each collection, is not statistically significantly bettered by any other retrieval model and is only -1 3% on average less the best results on the three collections The fusion of the other standard visual retrieval model, Manhattan distance, is the second best and according to our synthesis of the results Lidstone is the next best, making it the best language modelling-based fusion of the colour, edge and texture features It is on average 6 1% less than the best results on each collection and the fusion of no other retrieval model produces statistically significantly superior results on any of the TRECVid collections The fusion of Jelinek-Mercer results is a disappointing 12 5% on average lower than the best fusion result on the three collections but only 0 7 retrieval models are on average statistically significantly better than it – only Jensen-Shannon and Manhattan distance's fused results on TRECVid 2003 are statistically significantly better than it We believe that this may indicate a certain strength in this language model that is not indicated by the MAP statistic

In summary, the unbiased *Vis-CombWtScore* fusion of the colour, edge and texture results leads to better retrieval results than for a single feature on the TRECVid 2003 (+3 2% on average) and TRECVid 2004 (+28 5% on average) collections, while on the TRECVid 2002 collection only *Vis-CombJointPr* improves on the single best feature (+3 7% on average) The *Vis-CombWtScore* fusion method is the best overall fusion method for combining visual features according to this study The weights for the different collections are quite inconsistent and a non-weighted CombSUM fusion may be better since the somewhat similar *Vis-CombJointPr* (sum log-likelihood scores without normalisation) performs reasonably on the three collections We investigated CombSUM in this context for fusing the Jelinek-Mercer language model's colour, edge and texture results and found that while it is worse for TRECVid 2002, it is overall better than the weighted variant *Vis-CombWtScore* when considering the results across all three TRECVid collections (Mc Donald and Smeaton, 2005) We have not, as yet, evaluated the CombSUM fusion method on other retrieval models for this task, but we believe it likely that CombSUM

would also perform relatively well on these at least for TRECVid 2003 and 2004 collections This indicates that the tuning of feature fusion weights is unresolved and problematic

The best fusion results are overall achieved by combining the Jensen-Shannon distance's edge, colour and texture results and the best fusion of the language modelling results is for the Lidstone discounting language model, which is only 5% less on average than the Jensen-Shannon distance's result and for all three TRECVid collections no other retrieval model produces statistically significantly better results than it The Jelinek-Mercer language model performs relatively poorly with on average 7% lower MAP than the Jensen-Shannon result but shows some strength since it is only statistically significantly worse than the Jensen-Shannon distance and Manhattan distance's results on the TRECVid 2003 collection, and on all other collections no other retrieval model is statistically significantly better than it We have completed our analysis of video retrieval using a single visual example (image or video clip) and will now investigate the use of multiple visual examples for retrieving video shots in the next section

7.3 Experiments with multiple visual examples

In this section we investigate fusion methods for combining the results of multiple visual examples We compare the fusion methods *VisExs-CombJointPr* (sum log-likelihoods), *VisExs-CombRank* (Borda count), *VisExs-CombScore* (traditional CombSUM), *VisExs-CombMaxRank* (round-robin) and *VisExs-CombMaxScore*, which we previously described in Chapter 4, for combining the visual examples of each TRECVid topic and therefore the reported results represent the performance of a purely visual approach to the TRECVid search tasks

The *VisExs-CombJointPr* fusion method combines the previously presented *Vis-CombJointPr* results to create a simple generative model, while the other fusion methods combine the *Vis-CombWtScore* results, which as we have shown in the previous section is in general the best fusion method for combining multiple visual features

We present the average results across retrieval models for combining the visual examples in Table 21 for the TRECVid search tasks The *VisExs-CombMaxScore* and *VisExs-CombMaxRank* fusion methods produce the best average results for all three TRECVid collections For TRECVid 2003 and TRECVid 2004 the results for the *VisExs-CombMaxScore* fusion method are minutely better than the *VisExs-CombMaxRank* fusion method for both the unbiased and optimised results On TRECVid 2002 the unbiased results of *VisExs-CombMaxRank* are slightly better on average than those of *VisExs-CombMaxScore* and vice-versa for the optimised results The next best fusion method, *VisExs-CombScore*, combines the visual examples using average normalised score and is better than combining by normalised rank, *VisExs-CombRank*, on average for the TRECVid 2002 and TRECVid 2004 collections On TRECVid 2003 *VisExs-CombRank* is on average slightly better *VisExs-CombScore* The probability-based fusion method *VisExs-CombJointPr* performs on average very poorly for TRECVid 2002 and TRECVid 2004 and is worse than separate visual retrieval of each topic's visual examples, which is the *Vis-CombJointPr* result for this fusion method, by -58 7% and -67 9% on these two collections Its results are better on TRECVid 2003 where it achieves similar results to *VisExs-CombRank* and *VisExs-CombScore*

Table 21: Comparison of the average results across retrieval models for combining the visual examples using the *VisExs-CombJointPr*, *VisExs-CombRank*, *VisExs-CombScore*, *VisExs-CombMaxRank* and VisExs-CombMaxScore fusion methods on the (a) TRECVid 2002, (b) TRECVid 2003 and (c) TRECVid 2004 collections. The percentage improvement is relative to the mean of the individual visual examples' retrieval results *Vis-CombWtScore*, which is also shown here for comparison.

(a) TRECVid 2002 average results across retrieval models.

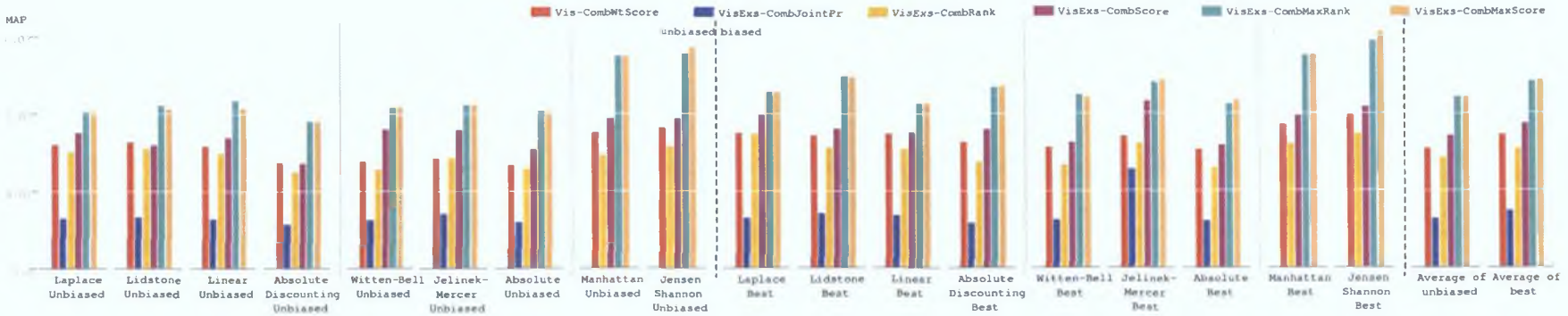
TRECVid 2002	*MAP*	*P@10*	*P@30*	*P@100*	*Impr.*
Vis-CombWtScore	**.0155**	.054	.033	.020	
Optimised	(**.0172**	.047	.030	.019)	
VisExs-CombJointPr	**.0064**	.026	.022	.015	-58.7%
Optimised	**.0073**	.026	.020	.015	-57.6%
VisExs-CombRank	**.0143**	.047	.042	.027	-7.1%
Optimised	(**.0153**	.048	.034	.025)	-11.0%
VisExs-CombScore	**.0171**	.062	.040	.026	+9.7%
Optimised	(**.0187**	.059	.037	.024)	+8.7%
VisExs-CombMaxRank	**.0222**	.082	.054	.028	+43.2%
Optimised	(**.0241**	.071	.043	.025)	+40.1%
VisExs-CombMaxScore	**.0221**	.082	.054	.028	+42.6%
Optimised	(**.0243**	.071	.043	.025)	+41.3%

(b) TRECVid 2003 average results across retrieval models.

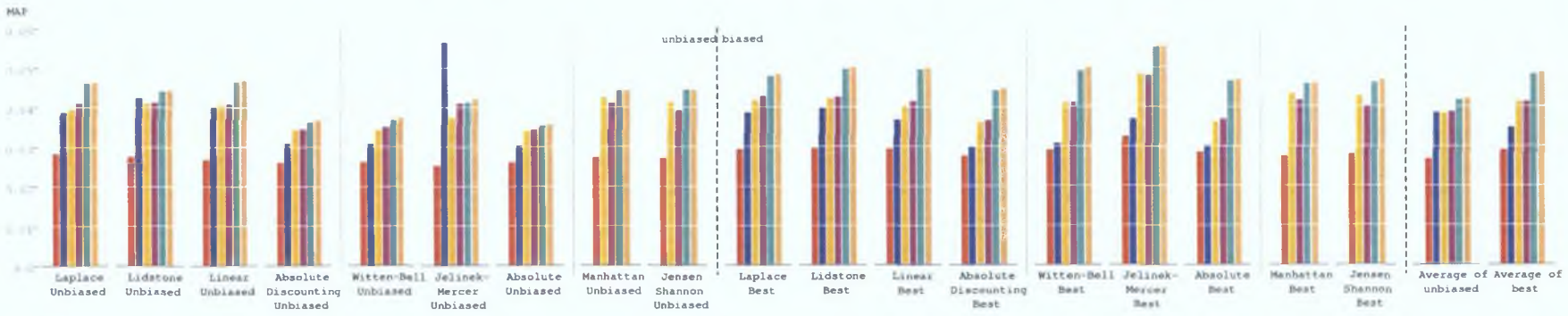
TRECVid 2003	*MAP*	*P@10*	*P@30*	*P@100*	*Impr.*
Vis-CombWtScore	**.0267**	.089	.072	.043	
Optimised	(**.0290**	.093	.076	.047)	
VisExs-CombJointPr	**.0385**	.077	.086	.057	+44.2%
Optimised	**.0347**	.073	.084	.054	19.7%
VisExs-CombRank	**.0383**	.096	.084	.060	+43.4%
Optimised	(**.0412**	.103	.091	.062)	+42.1%
VisExs-CombScore	**.0387**	.087	.090	.060	+44.9%
Optimised	(**.0412**	.092	.097	.061)	+42.1%
VisExs-CombMaxRank	**.0416**	.108	.092	.057	+55.8%
Optimised	(**.0482**	.120	.091	.062)	+66.2%
VisExs-CombMaxScore	**.0420**	.108	.092	.057	+57.3%
Optimised	(**.0486**	.120	.092	.062)	+67.6%

(c) TRECVid 2004 average results across retrieval models.

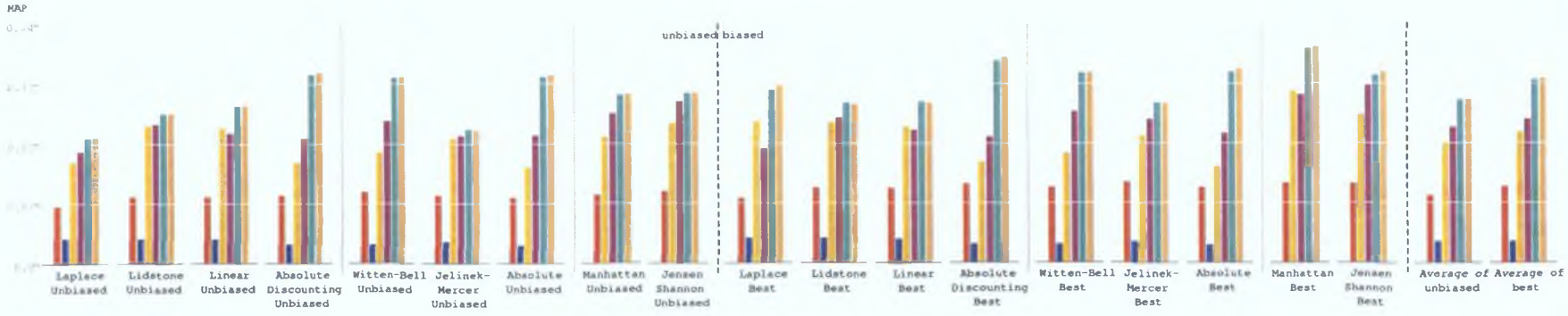
TRECVid 2004	*MAP*	*P@10*	*P@30*	*P@100*	*Impr.*
Vis-CombWtScore	**.0112**	.041	.028	.016	
Optimised	**.0127**	.042	.029	.017	
VisExs-CombJointPr	**.0036**	.030	.022	.013	-67.9%
VisExs-CombRank	**.0199**	.057	.048	.034	+77.7%
Optimised	(**.0218**	.056	.051	.036)	+71.7%
VisExs-CombScore	**.0226**	.073	.065	.034	+101.8%
Optimised	(**.0239**	.068	.064	.035)	+88.2%
VisExs-CombMaxRank	**.0239**	.094	.058	.033	+113.4%
Optimised	(**.0305**	.113	.070	.036)	+140.2%
VisExs-CombMaxScore	**.0273**	.100	.063	.034	+143.8%
Optimised	(**.0307**	.114	.070	.035)	+141.7%



(a) TRECVid 2002



(b) TRECVid 2003



(c) TRECVid 2004

Figure 59: Comparison of the fusion methods *VisExs-CombJointPr*, *VisExs-CombRank*, *VisExs-CombScore*, *VisExs-CombMaxRank* and *VisExs-CombMaxScore* for combining the visual examples' retrieval results for the different retrieval models on the (a) TRECVid 2002, (b) TRECVid 2003 and (c) TRECVid 2004 collections. The mean of the individual visual examples' retrieval results *Vis-CombWtScore* are also shown for comparison.

We compare the fusion results for combining the different retrieval model's visual results in Figure 59 for the TRECVid collections These results show that the *VisExs-CombMaxScore* fusion method and the *VisExs-CombMaxRank* fusion method are consistently very close for the different retrieval models The performance of the *VisExs-CombScore* and *VisExs-CombRank* fusion methods are also tied together with the combination by average score being favoured in most cases The *VisExs-CombJointPr* fusion method achieves very poor results for all language models on TRECVid 2002 and 2004 collections It is also the worst fusion approach for most of the language models on TRECVid 2003 except Lidstone and Jelmek-Mercer smoothed language models The surprisingly good performance on the Jelinek-Mercer language model on this collection is perhaps an outlier since it is not repeated on any other collection

Overall the results indicate that the combination of visual examples by maximum score or maximum rank (round-robin) is superior to the other fusion methods We compare the *VisExs-CombMaxScore* fusion method with the other fusion methods for the different retrieval models in Tables 65, 66 and 67 (appendix pp 270, 271 and 272) for TRECVid 2002, TRECVid 2003 and TRECVid 2004

The results for all language models on the TRECVid 2002 and 2003 collections show a large benefit, greater than 199%, for the *VisExs-CombMaxScore* fusion method compared to the *VisExs-CombJointPr* fusion method These large improvements are statistically significant for all language models on the TRECVid 2002 collection but for only two language models on TRECVid 2004, while the more modest improvements on the TRECVid 2003 collection are mostly statistically significant The relationship between the *VisExs-CombMaxScore* fusion method and other fusion methods is not as clear cut

The *VisExs-CombMaxScore* fusion method is better than combining by averages, such as with the *VisExs-CombRank* or *VisExs-CombScore* fusion methods, for all retrieval models on all tested TRECVid collections While very few of the positive improvements of *VisExs-CombMaxScore* on these collections are statistically significant, we believe that the trend shows that combining multiple visual examples by maximum score (or maximum rank) is better than by average rank or average score

The *VisExs-CombMaxScore* fusion method is better than the *VisExs-CombMaxRank* fusion method for all retrieval models on the TRECVid 2003 collection and all but the Lidstone and Jelinek-Mercer language models on the TRECVid 2004 collection, though none of these tiny 1% or 2% improvements are statistically significant (see Tables 66 and 67 in appendix pp 271 and 272) The two unbiased negative results for *VisExs-CombMaxScore* on the TRECVid 2004 collection show a statistically significant decease of -0 1 and -0 8 in the MAP for using the *VisExs-CombMaxScore* fusion method compared to *VisExs-CombMaxRank* fusion method The results for TRECVid 2002 in Table 65 (appendix page 270) are more mixed but overall favour the *VisExs-CombMaxRank* fusion method with about half of the unbiased combinations of retrieval models achieving statistically significantly better results for the *VisExs-CombMaxRank* fusion method compared to *VisExs-CombMaxScore* fusion method The other results on TRECVid 2002 that favour the *VisExs-CombMaxScore* fusion method over *VisExs-CombMaxRank* fusion method are not statistically significant We believe that while the *VisExs-CombMaxScore* fusion method is on average slightly better than the *VisExs-CombMaxRank* retrieval model for most retrieval models on the three collections, the statistically significant counter examples indicate

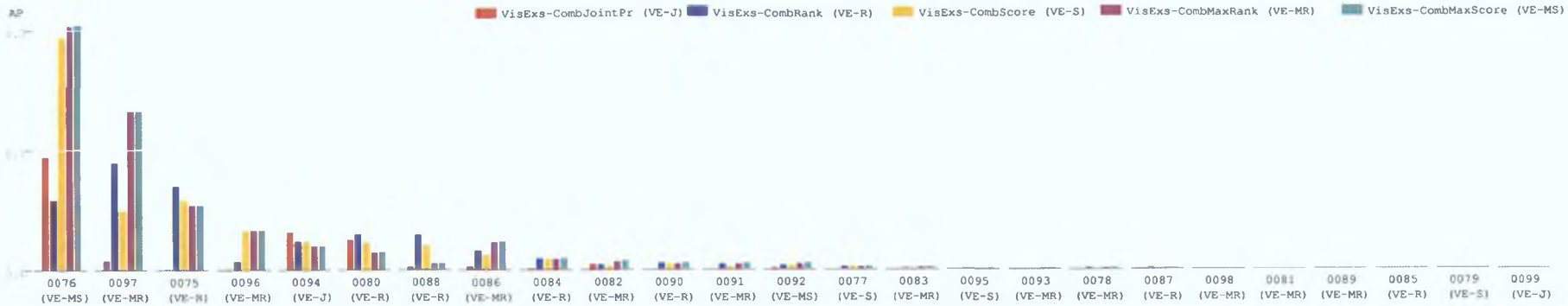
that the *VisExs-CombMaxRank* (round-robin) is superior and may have more consistent results across topics for combining visual examples

We compare the fusion methods for combining the Jelinek-Mercer language model's visual examples on the official TRECVid topics in Figure 60 This figure indicates that the visual-only results are very poor for many of the TRECVid topics, with the performance of all fusion methods in terms of average precision being almost non-existent for 6 topics in the TRECVid 2002 collection, 7 topics in the TRECVid 2003 collection and 9 topics in the TRECVid 2004 collection It is really only the top ten topics on the TRECVid 2002 and TRECVid 2003 collections and the top 7 topics on TRECVid 2004 that have a positive response to the visual-only query We can see that ranking by maximum score or maximum rank overall performs quite well for these top topics and that combining using the joint probability *VisExs-CombJointPr* fusion method performs better on the TRECVid 2003 collection than for the topics of the other collection Overall, visual-only searching is a poor strategy for the majority of the TRECVid topics

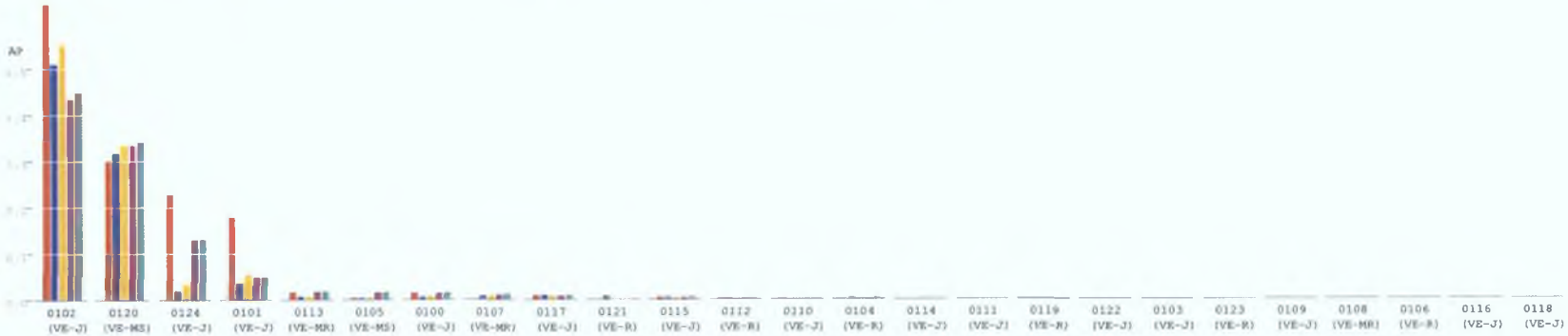
We will now compare our results for combining visual examples with the results for using the single best visual example for each topic We have shown previously in this section that the combination of visual examples using the *VisExs-CombMaxScore* fusion method improves on the mean performance for performing separate visual searches, as in the *Vis-CombWtScore* results (see Figure 59) In Figure 61 we compare the fusion of visual examples with using the single best image for each topic We can see that using the single best image for the colour-only or for the combined colour, edge and texture (*Vis-CombWtScore*) searching is better than using the combined visual examples The difference between the single best image searching and fused visual examples is greatest for TRECVid 2004 It is worth bearing in mind that there is no reason to believe that a user will know beforehand which topic image will perform best unless they are very familiar with the search collection We believe that user-based experiments would provide more insight into this issue and we also intend to investigate the combination of the best two and the best three topic examples in future experiments in order to see if the fusion methods can improve on retrieval that uses the best single image

We will now discuss the performance of the *VisExs-CombWtScore* fusion method in terms of the underlying visual retrieval models (see Tables 68, 69 and 70 in appendix pp 273, 273 and 274) Our best visual-only fusion results are generated using the Jensen-Shannon retrieval model for TRECVid 2002 (MAP 0 0286), the Linear language model for TRECVid 2003 (MAP 0 0468) and the Absolute discounting language model for TRECVid 2004 (MAP 0 0319) For TRECVid 2002 the best language model-based fusion combines the Jelinek-Mercer language model's visual example's results (MAP 0 0212), which is 26% lower than the Jensen-Shannon distance's fusion results, while the Jensen-Shannon distance's fusion results are the best standard visual retrieval results for the TRECVid 2003 and TRECVid 2004 collections where it is respectively 5 1% and 10 3% less than the best language modelling results

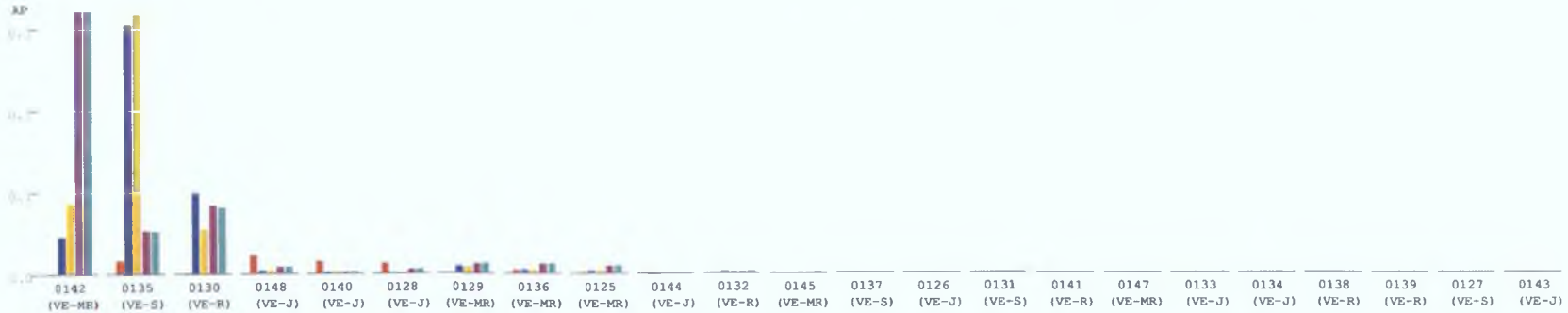
We present our synthesis of these results for the three TRECVid collections in Table 22 The main difference in this synthesis for the fusion of multiple visual examples and the previous synthesis for separate visual examples (fusion of visual features) in Table 20 is that the percentage difference between the best results (see the "%Dif" column) is in general doubled indicating that the results are more spread out in terms of the MAP measures Another difference is that



(a) TRECVid 2002

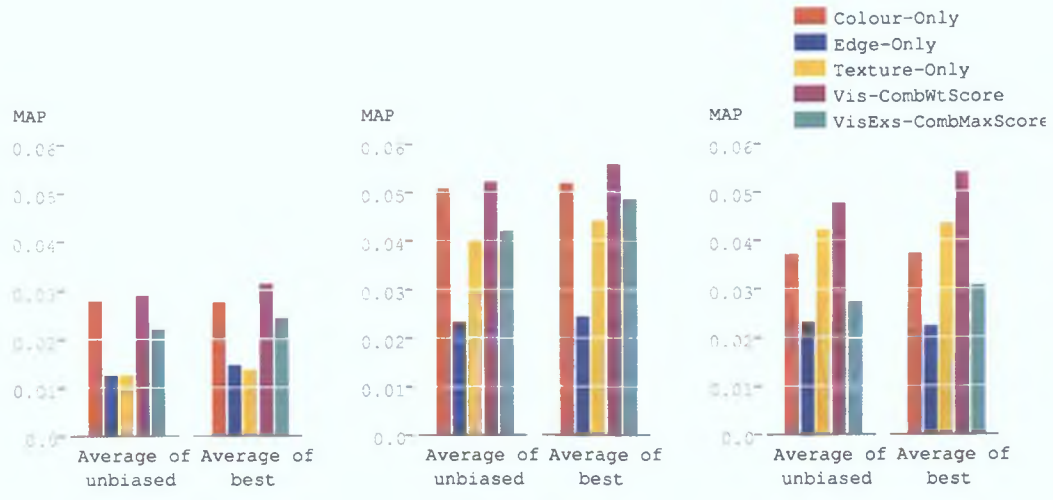


(b) TRECVid 2003



(c) TRECVid 2004

Figure 60: Comparison of the fusion methods *VisExs-CombJointPr*, *VisExs-CombRank*, *VisExs-CombScore*, *VisExs-CombMaxRank* and *VisExs-CombMaxScore* for combining the Jelinek-Mercer language model's results for the visual examples' results in each topic on the (a) TRECVid 2002, (b) TRECVid 2003 and (c) TRECVid 2004 collections.



(a) TRECVid 2002 (b) TRECVid 2003 (c) TRECVid 2004

Figure 61: Comparison of the fusion of all visual examples using *VisExs-CombMaxScore* with results for the single best image per topic for colour-only, texture-only, edge-only and combined-visual (*Vis-CombWtScore*) results.

Table 22: Comparison of unbiased retrieval models for the TRECVid 2002, 2003 and 2004 search tasks using the *VisExs-CombMaxScore* fusion method for combining colour, edge and texture results. %Dif = difference from best result (lower better). R = rank of retrieval models (lower better). S = number of statistically significantly better retrieval models (lower better).

	TRECVid'02		*TRECVid'03*		*TRECVid'04*		*Average*	
Ret. Method	*%Dif*	*R(S)*	*%Dif*	*R(S)*	*%Dif*	*R(S)*	*%Dif*	*R(S)*
Jensen-Shannon	0.0%	1(0)	-5.1%	3(0)	-10.3%	4(0)	**-5.1%**	**2.7(0.0)**
Manhattan	-3.8%	2(1)	-5.1%	4(0)	-11.0%	5(0)	**-6.6%**	**3.7(0.3)**
Linear	-27.6%	5(3)	0.0%	1(0)	-17.2%	6(1)	**-14.9%**	**4.0(1.3)**
Witten-Bell	-26.9%	4(0)	-20.3%	7(6)	-2.5%	3(0)	**-16.6%**	**4.7(2.0)**
Lidstone	-28.0%	6(3)	-5.3%	5(0)	-21.6%	9(3)	**-18.3%**	**6.7(2.0)**
Absolute Discounting	-33.6%	9(5)	-21.6%	8(6)	0.0%	1(0)	**-18.4%**	**6.0(3.7)**
Laplace	-29.0%	7(3)	-1.1%	2(0)	-34.2%	7(1)	**-21.4%**	**5.3(1.3)**
Jelinek-Mercer	-25.9%	3(0)	-10.3%	6(0)	-30.4%	8(0)	**-22.2%**	**5.7(0.0)**
Absolute	-29.4%	8(1)	-23.5%	9(6)	-1.3%	2(0)	**-26.5%**	**6.3(2.3)**

Lidstone and Laplace language models have been demoted by two places in our synthesised rankings of retrieval models We believe that our synthesised results are reasonably consistent for the two sets of experiments and that overall performance based on single visual example retrieval is a good indicator for the relative performance of the retrieval using multiple visual examples

The Jensen-Shannon distance model is the best retrieval model It is only 5 1% on average lower than the best results of the retrieval models on each collection, is on average the 2 7-th ranked retrieval model on each TRECVid collection, and is again not statistically significantly bettered by any other retrieval model Manhattan distance is again the second best retrieval model performing slightly worse than Jensen-Shannon distance with on average 6 6% worse MAP than the best results on all three collections Since Lidstone was demoted two places in our ranking of retrieval models, Linear is now our best language model result with on average 14 9% worse results than the best retrieval model on each collection and is on average the 4th retrieval model in terms of its MAP for each TRECVid collection Jelinek-Mercer is also again the second worst retrieval model in terms of MAP, which is 22 2% worse than the best MAP for each TRECVid collection but interestingly there is no statistically significantly better retrieval models than Jelinek-Mercer on any of the collections

We have therefore found that the Jelinek-Mercer smoothed language model has equivalent performance compared to the best standard visual retrieval model in terms of statistical significance for TRECVid visual-only multi-feature multi-example searching We have shown this previously for the regional edge and regional texture feature in the preceding chapter, where Jelinek-Mercer was overall better than Jensen-Shannon but not statistically significantly so We could possibly get better MAP for multi-feature multi-example search by fusing the Lidstone regional colour results with the Jelinek-Mercer edge and texture results but we have not investigated this and there is likely to be only a slight difference in the results The results for these discrete visual language modelling approaches are better than the best DCT GMM approach for TRECVid 2003 and 2004 The best DCT GMM approaches were static model for TRECVid 2002 and dynamic model for TRECVid 2003 and 2004 (see Table 2 on page 79) The Jelinek-Mercer language model achieves 26 1% worse MAP on TRECVid 2002 but 35 5% and 122 0% better MAP on the TRECVid 2003 and 2004 collections relative to the best DCT GMM approach (see Tables 68, 69 and 70 in appendix pp 273, 273 and 274) The Jensen-Shannon distance does better achieving equivalent results on TRECVid 2002 and 43 2% and 186 0% better results on TRECVid 2003 and 2004 respectively relative to the best DCT GMM query-likelihood approach

We have completed our fully automatic experiments for the visual-only retrieval of video shots within the TRECVid framework In summary, the *VisExs-CombMaxScore* and *VisExs-CombMaxRank* fusion methods are the best fusion methods for combining the retrieval results of multiple visual examples The Jensen-Shannon distance's results were the best overall for the *VisExs-CombMaxScore* fusion of multiple visual examples, while the Linear language model was the best language modelling result in terms of average MAP with on average 10% lower MAP than the Jensen-Shannon distance's results The Jelinek-Mercer language model, which has an average retrieval rank of 5 7 compared to other retrieval models for the three collections and is on average 17% worse than the Jensen-Shannon distance's results in terms of MAP, is interestingly not statistically significantly bettered by another retrieval model on any of the tested TRECVid collections We will use the *VisExs-CombMaxScore* fusion method as the basis for our text and

visual fusion experiments in the next section as this is the best fusion method for the majority of the unbiased results and its use will allow us to further explore the normalised score-based fusion methods in the context of combining text and visual results However, it should be borne in mind that the *VisExs-CombMaxRank* fusion method is probably superior since some of our results have small but statistically significant improvements for the *VisExs-CombMaxRank* fusion method compared to the *VisExs-CombMaxScore* fusion method

7.4 Experiments with multiple modalities

In this section we present experiments that combine the text-based *shot+adj+video* hierarchical language model's results with the visual-based results that were achieved by combining multiple visual examples using the *VisExs-CombMaxScore* fusion method Specifically, we will compare the fusion methods *TextVis-CombWtScore* (weighted sum of normalised scores), *TextVis-CombWtRank* (weighted Borda count) and *TextVis-CombJointPr* (sum log-likelihoods), which we previously described in Chapter 4

We present the average results of the fusion methods across the different retrieval models in Table 23 The *TextVis-CombJointPr* probability-based text and visual fusion method produces on average far worse results than for text alone on all three TRECVid collections It achieves more that 70% poorer results on the TRECVid collections than for text alone The *TextVis-CombWtScore* fusion method produces overall the best unbiased results on all three collections, achieving an increase of 0 4% on TRECVid 2002, 7 7% on TRECVid 2003 and 9 7% on TRECVid 2004 compared to the text-only results These results are quite encouraging as they indicate that a completely automatic combination of text and visual examples produces a positive improvement on all three collections For the TRECVid 2003 and TRECVid 2004 collections the precision at 10, 30 and 100 document cut-offs is also improved for the *TextVis-CombWtScore* fusion method compared to the text-only results This indicates that a user's initial ranking from the default TRECVid query (text + visual examples) will produce on average better results for their top ranking documents than for text alone The *TextVis-CombWtRank* fusion method was unable to combine the text and visual results on the TRECVid 2002 collection – the optimised weight for the visual results was zero for most retrieval models on this collection However its optimised results on TRECVid 2003 are better than for *TextVis-CombWtScore*, showing an average increase of 16 2% compared to *TextVis-CombWtScore* fusion method's 10 9% increase on the text-only results, but for the TRECVid 2004 collection its unbiased and optimised results are on average lower than the *TextVis-CombWtScore* fusion method's results

In Figure 62 we compare the fusion methods for combining the text and visual results of the different visual retrieval models on the three TRECVid collections The results for fusing the different visual retrieval models are essentially consistent within each TRECVid collection On the TRECVid 2002 collection the *TextVis-CombWtScore* fusion method is the only fusion method to produce better unbiased fusion results for any of the retrieval models and it also produces the best improvement of all fusion methods for all the optimised results on this collection (see Figure 62(a)) On the TRECVid 2003 collection the *TextVis-CombWtScore* fusion method produces better unbiased fusion results than for the text-only and other fusion methods, while the *TextVis-CombWtRank* fusion method produces the best results for the optimised fusion results for all visual retrieval models (see Figure 62(b)) On the TRECVid 2004 collection the

Table 23: Comparison of the average results across retrieval models for combining the *shot+adj+video* interpolated text language model's results with the combined visual examples (*VisExs-CombMaxScore*) results of the different retrieval models using the *TextVis-CombJointPr*, *TextVis-CombWtRank* and *TextVis-CombWtScore* fusion methods on the (a) TRECVid 2002, (b) TRECVid 2003 and (c) TRECVid 2004 collections. The improvement percentage is relative to the text-only results. The text-only and the visual-only, *VisExs-CombMaxScore*, results are displayed for comparison.

(a) TRECVid 2002 average results across retrieval models.

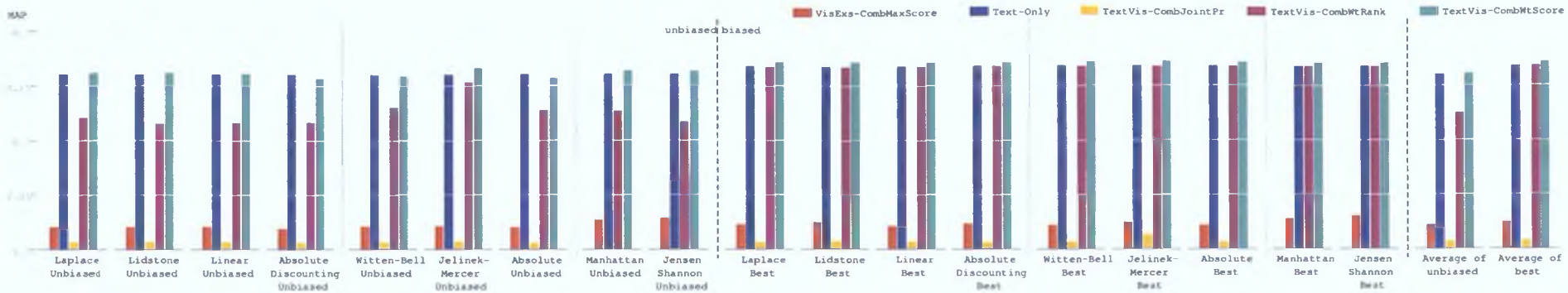
TRECVid 2002	*MAP*	*P@10*	*P@30*	*P@100*	*Impr.*
Text-Only	**.1605**	.264	.175	.117	
Optimised	(**.1677**	.272	.175	.118)	
VisExs-CombMaxScore	**.0221**	.082	.054	.028	
Optimised	(**.0243**	.071	.043	.025)	
TextVis-CombJointPr	**.0064**	.026	.022	.015	-96.0%
Optimised	**.0073**	.026	.020	.015	-95.6%
TextVis-CombWtRank	**.1250**	.234	.184	.119	-22.1%
Optimised	(**.1677**	.272	.175	.118)	+0.0%
TextVis-CombWtScore	**.1612**	.263	.186	.122	+0.4%
Optimised	(**.1713**	.270	.177	.121)	+2.2%

(b) TRECVid 2003 average results across retrieval models.

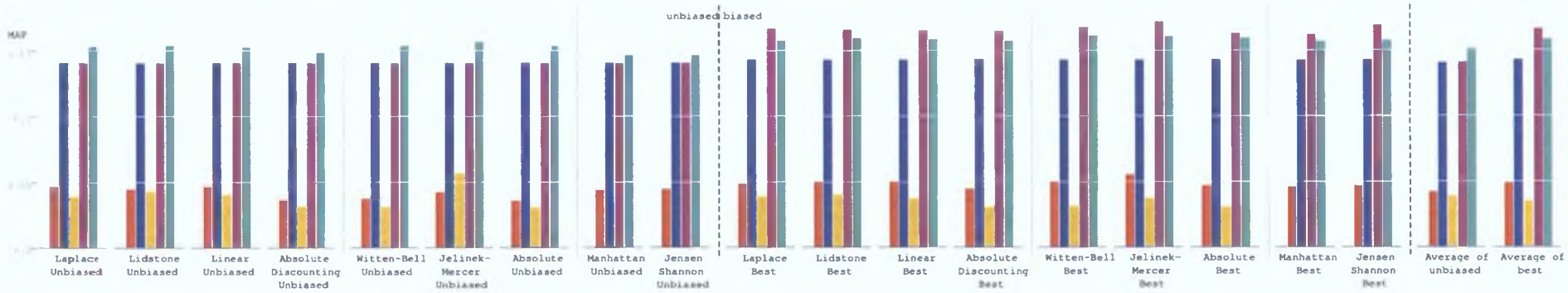
TRECVid 2003	*MAP*	*P@10*	*P@30*	*P@100*	*Impr.*
Text-Only	**.1405**	.252	.176	.113	
Optimised	(**.1429**	.240	.179	.115)	
VisExs-CombMaxScore	**.0420**	.108	.092	.057	
Optimised	(**.0486**	.120	.092	.062)	
TextVis-CombJointPr	**.0385**	.077	.086	.057	-72.6%
Optimised	**.0347**	.073	.084	.054	-75.7%
TextVis-CombWtRank	**.1405**	.252	.176	.113	+0.0%
Optimised	(**.1661**	.303	.202	.119)	+16.2%
TextVis-CombWtScore	**.1514**	.266	.187	.117	+7.7%
Optimised	(**.1585**	.279	.196	.121)	+10.9%

(c) TRECVid 2004 average results across retrieval models.

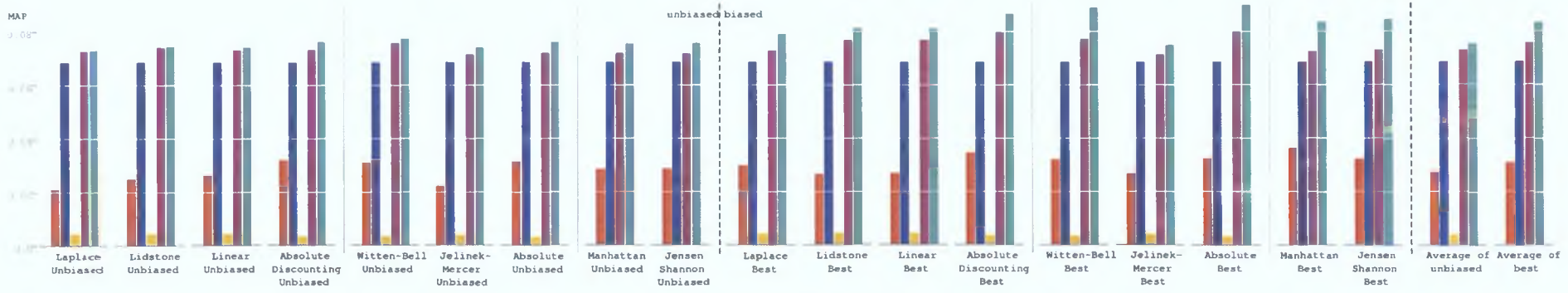
TRECVid 2004	*MAP*	*P@10*	*P@30*	*P@100*	*Impr.*
Text-Only	**.0686**	.209	.143	.091	
VisExs-CombMaxScore	**.0242**	.094	.058	.033	
Optimised	(**.0307**	.114	.070	.035)	
TextVis-CombJointPr	**.0036**	.030	.022	.013	-94.8%
TextVis-CombWtRank	**.0730**	.252	.159	.093	+6.4%
Optimised	(**.0756**	.252	.158	.093)	+10.3%
TextVis-CombWtScore	**.0752**	.219	.157	.100	+9.7%
Optimised	(**.0833**	.258	.180	.096)	+21.4%



(a) TRECVid 2002



(b) TRECVid 2003



(c) TRECVid 2004

Figure 62: Comparison of the fusion methods *TextVis-CombJointPr*, *TextVis-CombWtRank* and *TextVis-CombWtScore* for combining the text and visual retrieval results of the different retrieval models on the (a) TRECVid 2002, (b) TRECVid 2003 and (c) TRECVid 2004 collections. The *text-only* and visual-only *VisExs-CombMaxScore* retrieval results are shown for comparison.

TextVis-CombWtScore fusion method produces the best unbiased and optimised results for all retrieval models (see Figure 62(c)) The *TextVis-CombJointPr* fusion method causes a drastic decrease in performance of the fusion of every retrieval model's results on the three collections by reducing their performance to that of the visual-only results *VisExs-CombJointPr* The reason this occurs is because the visual samples are so large in terms of predicted events (all pixels for colour and edge features in the example images and all 8x8 image blocks for DCT feature in the example images) that they overwhelm the text probabilities from the very small query text when combined using joint probability

We compare the *TextVis-CombWtScore* fusion method with the other fusion methods for all retrieval models in Tables 71, 72 and 73 (appendix pp 275, 276 and 277) for the TRECVid 2002, 2003 and 2004 collections We will now discuss these results in turn for each of the TRECVid collections

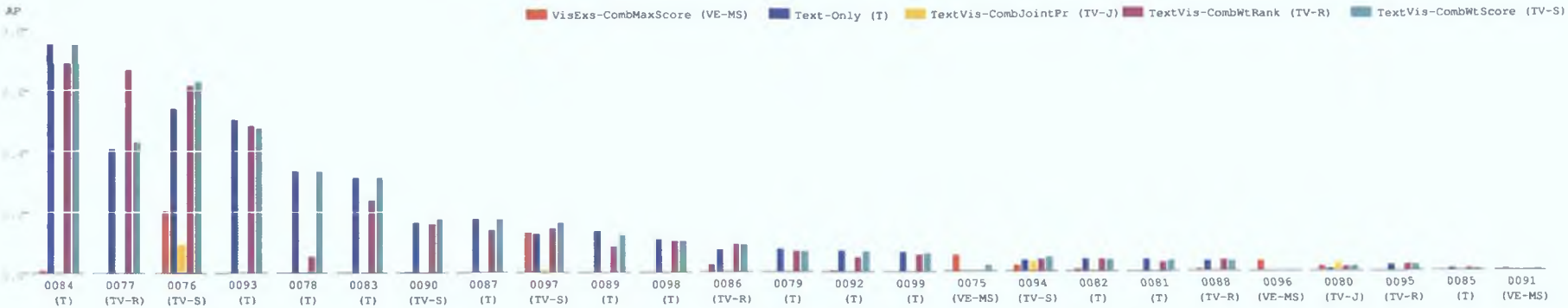
The results for the TRECVid 2002 collection in Table 71 (appendix page 275) show that the *TextVis-CombWtScore* fusion method produces statistically significantly better fusion results than the *TextVis-CombJointPr* fusion method for all visual retrieval models and it also has statistically significantly better unbiased results than the *TextVis-CombWtRank* fusion method for all but one (Absolute interpolation) of the visual retrieval models Most of the unbiased *TextVis-CombWtScore* results are better than the text-only results though none of these improvements are statistically significant The unbiased fusion of the Jelinek-Mercer visual results with the text results achieves the best performance with a MAP of 0 1663, a 3 6% improvement on the text-only results The best optimised results are also from the fusion of Jelinek-Mercer's visual results, which achieves a MAP of 0 1720, a 2 6% improvement on the optimised text-only results The fusion of the Manhattan distance's visual results achieves the second best performance with an unbiased MAP of 0 1637, a 1 9% improvement on text-only results, and with an optimised MAP of 0 1712, a 2 1% improvement on optimised text-only results Jensen-Shannon distance achieves the same optimised result as Manhattan distance and achieves only slightly less than Manhattan distance, 0 1633 compared to 0 1637, for the unbiased results On this collection the optimised weights for the visual results vary from 0 1 to 0 25 for combining the different visual retrieval model's results with text-only results

The results for the TRECVid 2003 collection in Table 72 (appendix page 276) show that the *TextVis-CombWtScore* fusion method is again statistically significantly better than all the *TextVis-CombJointPr* results The *TextVis-CombWtScore* fusion results are better than the text-only results for every retrieval model and the unbiased fusion results for Lidstone, Linear, Manhattan and Jensen-Shannon distance are actually statistically significantly better than the text-only results The *TextVis-CombWtScore* fusion method performs better than the *TextVis-CombWtRank* for all unbiased results, however we cannot take much cognisance of this observation due to the poor performance of the *TextVis-CombWtRank* fusion method on TRECVid 2002 where the optimised weight for the visual results was zero for many of the retrieval models The *TextVis-CombWtRank* fusion method performs better than the *TextVis-CombWtScore* method for all optimised results on the TRECVid 2003 collection but none of these improvements are statistically significant Similar to the TRECVid 2002 collection, the *TextVis-CombWtScore* combination of the Jelinek-Mercer's visual results again achieves the highest unbiased result with a MAP of 0 1564, an 11 3% improvement on the text-only results The best optimised *TextVis-CombWtScore* fusion result is for combining the Witten-Bell visual results, which achieves a MAP

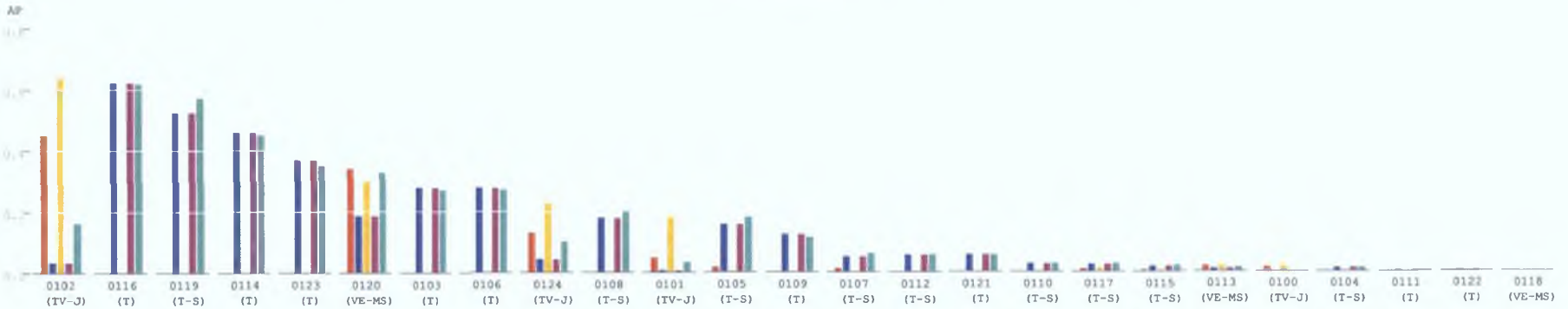
of 0 1611, a 12 7% improvement on the optimised text-only results The optimised Jelinek-Mercer fusion is not far behind and is the next best with a MAP of 0 1605 The fusion of Manhattan distance's and Jensen-Shannon distance's results both achieve the worst results, an unbiased MAP of 0 1462, which is a statistically significant 4% improvement on the text-only results All the optimised *TextVis-CombWtScore* fusions of text and visual results give text results a weight of 0 7 and visual results a weight of 0 3 on the TRECVid 2003 collection

The results for the TRECVid 2004 collection in Table 73 (appendix page 277) show that the *TextVis-CombWtScore* fusion method is statistically significantly better than the *TextVis-CombJointPr* fusion method for all retrieval models The *TextVis-CombWtScore* fusion method is also better than the *TextVis-CombWtRank* fusion method for all optimised and unbiased combinations though none of these improvements are statistically significant The fusion of the Witten-Bell language model's results achieve the best unbiased results with a MAP of 0 0775, a 13 1% improvement on the text-only results and the best optimised fusion was for the Absolute interpolation results with a MAP of 0 0899, a 31 0% improvement on the text-only results The fusion of Jensen-Shannon distance's results produced the best results for the standard visual retrieval models, achieving an unbiased MAP of 0 0758 (+10 5%) and an optimised MAP of 0 844 (+23 1%) While the fusion of the Jelinek-Mercer language model's results, which was the best result for the TRECVid 2002 and TRECVid 2003 collections, has the second worst unbiased results with a MAP of 0 0741, an 8 1% improvement on text-only results, and it also has the worst optimised fusion with a MAP of 0 0750, a 9 3% improvement on the text-only results Most of the optimised fusion methods give the visual results a weight of 0 45 and the text results a weight of 0 55, which indicates the surprisingly high importance of the visual results in improving the optimised rankings on the TRECVid 2004 collection

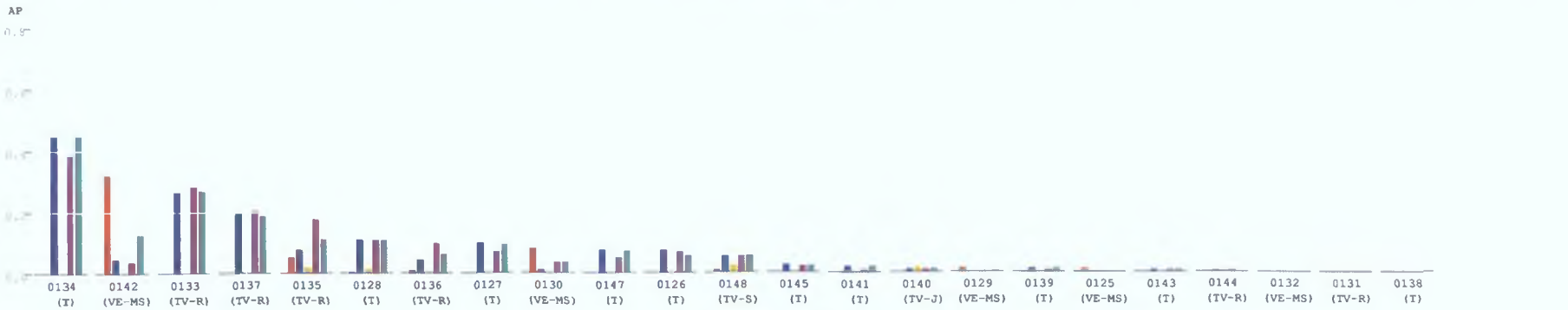
We will now discuss the topic results for the fusion of the text-only results with the visual-only Jelinek-Mercer language model's results as shown in Figure 63 for the three TRECVid collections We further present some comparisons between the different fusion methods for the Jelinek-Mercer's visual results in Table 24 Table 24(a) indicates that over 58 9% of the TRECVid topics are better suited to non-multimodal search such as either text-only (43 8% of topics) or visual-only (15 1% of topics) retrieval, while the other 41 1% of TRECVid topics are best suited to multimodal retrieval using either of the three fusion methods These multimodal-oriented topics perform best with the *TextVis-CombWtScore* (19 2% of topics) fusion method compared to the other fusion methods *TextVis-CombJointPr* (6 8% of topics) and *TextVis-CombWtRank* (13 7% of topics) We compare the fusion methods separately with the text-only results in Tables 24(b), 24(c) and 24(d) We see from these tables that the *TextVis-CombJointPr* fusion method is better for only 11% of topics, the *TextVis-CombWtScore* fusion method is better for 47% of topics and the *TextVis-CombWtRank* fusion method is better for 32% of topics compared to the text-only results This indicates that the *TextVis-CombWtScore* fusion method also produces a higher proportion of better topic results in comparison to the text-only results for the full set of TRECVid topics than the other fusion methods We directly compare the best fusion method *TextVis-CombWtScore* separately with the other two fusion methods for the full set of TRECVid topics in Table 24(e) and 24(f) We see from both tables that *TextVis-CombWtScore* is better for about 60% of topics when compared directly with the *TextVis-CombWtRank* fusion method and about 90% of topics when compared with *TextVis-CombJointPr* fusion method The *TextVis-CombWtScore* is therefore best compared to other tested fusion methods for the subset of multimodal oriented topics and for the full set of topics at least for fusing the Jelinek-Mercer



(a) TRECVid 2002



(b) TRECVid 2003



(c) TRECVid 2004

Figure 63: Comparison of the fusion methods *TextVis-CombJointPr*, *TextVis-CombWtRank* and *TextVis-CombWtScore* for combining the text results with the Jelinek-Mercer language model's visual results on the (a) TRECVid 2002, (b) TRECVid 2003 and (c) TRECVid 2004 collections. The text-only and visual-only results *VisExs-CombMaxScore* are shown for comparison.

Table 24 Comparison of the text-only, visual-only and fusion methods *TextVis-CombJointPr*, *TextVis-CombWtRank* and *TextVis-CombWtScore* for combining the text results with the Jelinek-Mercer language model's visual results on the TRECVid 2002, TRECVid 2003 and TRECVid 2004 collections

(a) Text-Only V Visual-Only V Fusion methods

	TV'02	TV'03	TV'04	Avg
Text-Only	13	9	10	43 8%
Visual-Only	3	3	5	15 1%
TextVis-CombWtScore	4	9	1	19 2%
TextVis-CombWtRank	4	0	6	13 7%
TextVis-CombJointPr	1	4	1	6 8%

(b) Text-Only V TextVis-CombJointPr

	TV'02	TV'03	TV'04	Avg
Text-Only	23	21	21	89%
TextVis-CombJointPr	2	4	2	11%

(c) Text-Only V TextVis-CombWtScore

	TV'02	TV'03	TV'04	Avg
Text-Only	16	10	13	53%
TextVis-CombWtScore	9	15	10	47%

(d) Text-Only V TextVis-CombWtRank

	TV'02	TV'03	TV'04	Avg
Text-Only	14	25	11	68%
TextVis-CombWtRank	11	0	12	32%

(e) TextVis-CombWtScore V TextVis-CombWtRank

	TV'02	TV'03	TV'04	Avg
TextVis-CombWtScore	15	15	13	59%
TextVis-CombWtRank	10	10	10	41%

(f) TextVis-CombWtScore V TextVis-CombJointPr

	TV'02	TV'03	TV'04	Avg
TextVis-CombWtScore	24	22	21	92%
TextVis-CombJointPr	1	3	2	8%

Table 25 Comparison of unbiased retrieval models for the TRECVid 2002, 2003 and 2004 search tasks using the *TextVis-CombMaxScore* fusion method for combining text and results %Dif = difference from best result (lower better) R = rank of retrieval models (lower better) S = number of statistically significantly better retrieval models (lower better)

	TRECVid'02		*TRECVid'03*		*TRECVid'04*		*Average*	
Ret Method	*%Dif*	*R(S)*	*%Dif*	*R(S)*	*%Dif*	*R(S)*	*%Dif*	*R(S)*
Jelinek-Mercer	0 0%	1(0)	0 0%	1(0)	-4 4%	8(1)	**-1 5%**	**3 3(0 3)**
Witten-Bell	-4 4%	7(0)	-1 9%	3(1)	0 0%	1(0)	**-2 1%**	**3 7(0 3)**
Lidstone	-2 8%	5(2)	-1 8%	2(0)	-3 9%	6(1)	**-2 8%**	**4 3(1 0)**
Absolute	-5 6%	8(4)	-2 0%	4(2)	-1 7%	3(1)	**-3 1%**	**5 0(2 3)**
Linear	-3 2%	6(5)	-2 6%	6(1)	-4 1%	7(1)	**-3 3%**	**6 3(2 3)**
Manhattan	-1 6%	2(0)	-6 5%	8(0)	-2 5%	5(0)	**-3 5%**	**5 0(0 0)**
Jensen-Shannon	-1 8%	3(0)	-6 5%	9(0)	-2 2%	4(0)	**-3 5%**	**5 3(0 0)**
Laplace	-2 7%	4(3)	-2 2%	5(1)	-5 9%	9(3)	**-3 6%**	**6 0(2 3)**
Absolute Discounting	-5 7%	9(5)	-5 2%	7(1)	-1 5%	2(1)	**-4 1%**	**6 0(2 3)**

language model's visual results

We will now discuss the performance of the fused text and visual results in terms of the underlying visual retrieval models, which we present in a synthesised form in Table 25 Statistical significance comparisons of the results for the three collections is available in Tables 74, 75 and 76 (appendix pp 278–279) The fused text and visual fusion results are very close on each collection – the worst fusion results are only respectively 5 7%, 6 5% and 5 9% lower than the best fusion results on the TRECVid 2002, 2003 and 2004 collections The fusion of visual Jelinek-Mercer language model's results achieves the best MAP on the TRECVid 2002 and TRECVid 2003 collections and in our synthesised results it is best retrieval model overall with an average rank of 3 3 and with on average only 0 3 statistically significantly better retrieval models above it on each of the TRECVid collections Only the visual Witten-Bell results on TRECVid 2004 collection is statistically significantly better This is in contrast to the fusion results for separate visual example retrieval (combined features) and the fusion results for multiple visual examples where the fusion of Jelinek-Mercer language model's results was the second worst overall retrieval model But for each of these results it had a very low number of statistically significantly better retrieval models above it and this may have indicated some latent consistency within its results that makes it better for combining with the better text results The Manhattan distance and Jensen-Shannon distance retrieval models are ranked below many of the language model's results at the sixth and seventh position in our synthesised results These results are only on average 3 5% lower than the best result on each collection and they are not statistically significantly bettered by any other retrieval model They are also on average only 2% lower than the Jelinek-Mercer result We believe that these results are just as good as the Jelinek-Mercer results even though their MAP is slightly lower The fusion of Witten-Bell language model's results performs quite strongly and achieves very close to the best Jelinek-Mercer result in terms of average rank, difference from best results and also has the same incidence rate of 0 3 for being statistically significantly bettered The other language models Laplace, Lidstone, Linear, Absolute interpolation and Absolute discounting are possibly weaker than the Jelinek-Mercer language model but these fusion results are so close that it would be unfair to make any strong claim

We have presented results for fully automatic retrieval experiments for the multimodal TRECVid topics We found that the *TextVis-CombWtScore* fusion method performed better than the other fusion methods, achieving on average better results on all three collections than for text alone The probability based *TextVis-CombJointPr* fusion method performed consistently poorly and reduced to the performance of the visual-only approach The *TextVis-CombWtRank* fusion method performed so poorly on the TRECVid 2002 collection that its unbiased results on TRECVid 2003 cannot be trusted as a true unbiased test of its performance We believe that the *TextVis-CombWtRank* fusion method is in general as good as the *TextVis-CombWtScore* fusion method This claim can be supported by looking at the optimised results on the TRECVid 2003 collection where it is superior to the *TextVis-CombWtScore* fusion method and by looking at its unbiased results on the TRECVid 2004 collection where it is only on average 3 1% lower than the *TextVis-CombWtScore* fusion method We found that the *TextVis-CombWtScore* fusion method achieved the best overall fusion results by using the Jelinek-Mercer language model's visual results The fusion results were overall very close and therefore it is hard to discriminate between visual language models and standard retrieval models

The three TRECVid collections indicate a trend in the optimised parameters of the *TextVis-CombWtScore* fusion method that the visual medium is more important for each subsequent year's TRECVid experiments For TRECVid 2002 the optimum weight for the visual results is between 0 1 and 0 25, for TRECVid 2003 it is 0 3 and for TRECVid 2004 it is 0 45 These results indicate that the formulation of TRECVid queries is progressively better each year at reducing the dominance of text retrieval in the experiments The more careful identification of potential topics without listening to the audio for TRECVid 2004 is probably one measure that accounts for some of the difference between TRECVid 2003 and 2004 fusion results Another trend that is important to recognise with automatic retrieval is that better visual results do not mean better text and visual fusion results The reason for this is that while visual results will improve some topics they will also degrade the performance of others The fusion performance may actually be greatly affected by how much it degrades high-performing text-only topics than poorly performing topics that do not contribute much to the MAP statistic This problem affects not only the unbiased results but also the optimised results due to the fact that we are trying to find a single weight that will give best performance across all topics

We will now consider two Oracle-based scenarios for combining the text and visual results, which will allow us to compare the text and visual retrieval models under ideal decision making situations where the Oracle always chooses the best of available choices for each topic In the first scenario, the *TextVisBoth-Oracle* selects whether the TRECVid topic will be text-only, visual-only or combined text and visual, while in the second scenario, the *TextVisComb-Oracle* selects the weights for combining text and visual results on a per-topic basis In both variation, of course, the Oracle selects the best of the available choices since it is omnipotent with regards to the outcome Both these variations are biased of course but as before we will present two sets of results, firstly for combining the unbiased visual retrieval models with the unbiased text retrieval models and secondly for combining the optimised visuals models with the optimised text models Remember in this thesis we are exploring fully automatic shot retrieval but in practice the retrieval system will be used by a user in an interactive search mode so the purpose of these scenarios is to investigate the potential performance our multimodal retrieval models could achieve in these two different user-interaction scenarios (i e users selecting query modes text, or visual or both text and visual for each TRECVid topic or users moving a slider to decide

how much importance to give to the text or visual modes for each TRECVid topic)

In a sense our first Oracle scenario is an optimised version of (Yan et al , 2004), which used a limited number of query-classes (finding persons, named objects, general objects, or scenes) for combining text and other retrieval modalities with pre-defined query-class weights, whereas here we assume that a user chooses correctly between three competing search strategies – text-only, visual-only or combined text and visual search for each TRECVid topic The second Oracle scenario can also be viewed as the upper bound performance for an automatic query-class dependent procedure that combines our default text and visual results for each TRECVid topic

We display the average improvement across visual retrieval models for the two Oracle scenarios in Table 26 For TRECVid 2002 the *TextVisBoth-Oracle* improves on the text-only results by only 6%, while the *TextVisComb-Oracle* improves on this result by only a further 3% There does not seem much justification in allowing the user to choose the weights for the text and visual fusion on the TRECVid 2002 collection as 3% is a negligible improvement for a scenario where the user must make the correct weight selection for each topic The *TextVisBoth-Oracle*'s improvement of 6% higher MAP than for text alone indicates only a small benefit of multimedia retrieval on this collection The results for TRECVid 2003 are far better with on average a 24 4% improvement for the *TextVisBoth-Oracle* compared to the text-only results, which is a 15 5% average improvement on the standard text and visual fusion and the *TextVisComb-Oracle* improves on this result by a further 4 6% The results for TRECVid 2004 are further proof of the usefulness of visual searching for the TRECVid topics The *TextVisBoth-Oracle* achieves on average +31 6% better results than for text alone, or 19 9% better than the default combination of text and visual results A further 9 3% on average is achievable through the correct manipulation of the fusion weights by the *TextVisComb-Oracle*

We compare the oracle scenarios *TextVisBoth-Oracle* and *TextVisComb-Oracle* with the visual-only, text-only and standard text and visual fusion method *TextVis-CombWtScore* in Figure 64 for the different retrieval model's results Unsurprisingly, the results for the two scenarios are quite consistent across retrieval models on each collection The performance of the Oracles' results are more directly related to the individual performance of the visual results than for the standard fusion models because the *TextVisBoth-Oracle* can ignore the visual results for a topic if its fusion would achieve lower results than for text alone and likewise it can ignore the text-based results if its fusion with the visual results would produce worse results than for visual alone Also, the *TextVisComb-Oracle* chooses the exact best fusion weight for combining each visual result with the text-based result for each topic and therefore gets the most out of the visual and text-based results

Overall the results for TRECVid 2003 and TRECVid 2004 are at least 24 4% improved compared to the text-only results when using the *TextVisBoth-Oracle* that simply chooses between text-only, visual-only and multimodal queries This result justifies including visual querying in many video databases that currently use text-only interfaces If we were to argue for allowing a user to choose the weight between text and visual components, which we're not, then it could possibly be justified from TRECVid 2004 results but we believe the other collections do not show enough of an improvement to warrant such an interface component The Oracle-based results have a more direct importance to our research in that these results for TRECVid 2003 and TRECVid 2004 indicate the need to further extend our work to allow for dynamic weighting in

Table 26: Comparison of the average results across retrieval models for combining the *Shot+Adj+Video* interpolated text language model's results with the combined visual examples (*VisExs-CombMaxScore*) results of the different retrieval models using the *TextVisBoth-Oracle* and *TextVisComb-Oracle* on the (a) TRECVid 2002, (b) TRECVid 2003 and (c) TRECVid 2004 collections. The improvement percentage is relative to the default text and visual fusion results. The text-only, visual-only, *VisExs-CombMaxScore* and default text and visual fusion (*TextVis-CombWtScore*) results are displayed for comparison.

(a) TRECVid 2002 average results across retrieval models.

TRECVid 2002	*MAP*	*P@10*	*P@30*	*P@100*	*Impr.Text*	*Impr.Both*	*Impr.Oracle*
Text-Only	**.1605**	.264	.175	.117			
Optimised	(**.1677**	.272	.175	.118)			
VisExs-CombMaxScore	**.0221**	.082	.054	.028			
Optimised	(**.0243**	.071	.043	.025)			
TextVis-CombWtScore	**.1612**	.263	.186	.122	+0.4%		
Optimised	(**.1713**	.270	.177	.121)	+2.2%		
TextVisBoth-Oracle	**.1701**	.283	.186	.123	+6.0%	+5.6%	
Optimised	(**.1819**	.291	.186	.127)	+8.5%	+6.2%	
TextVisComb-Oracle	**.1752**	.291	.188	.126	+9.1%	+8.7%	+3.0%
Optimised	(**.1829**	.294	.187	.128)	+9.1%	+6.7%	+0.5%

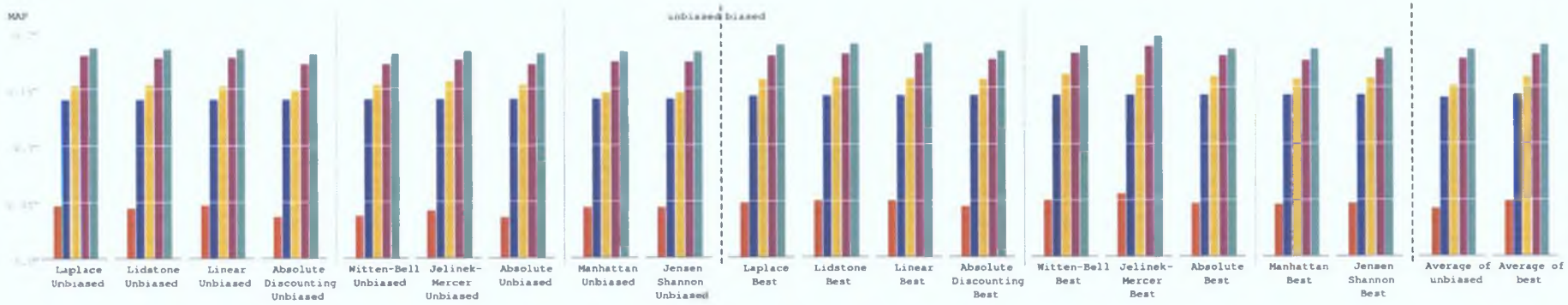
(b) TRECVid 2003 average results across retrieval models.

TRECVid 2003	*MAP*	*P@10*	*P@30*	*P@100*	*Impr.Text*	*Impr.Both*	*Impr.Oracle*
Text-Only	**.1405**	.252	.176	.113			
Optimised	(**.1429**	.240	.179	.115)			
VisExs-CombMaxScore	**.0420**	.108	.092	.057			
Optimised	(**.0486**	.120	.092	.062)			
TextVis-CombWtScore	**.1514**	.266	.187	.117	+7.7%		
Optimised	(**.1585**	.279	.196	.121)	+10.9%		
TextVisBoth-Oracle	**.1748**	.326	.231	.141	+24.4%	+15.5%	
Optimised	(**.1785**	.314	.233	.144)	+24.9%	+12.6%	
TextVisComb-Oracle	**.1828**	.334	.238	.143	+30.1%	+20.8%	+4.6%
Optimised	(**.1867**	.332	.240	.150)	+30.6%	+17.8%	+4.6%

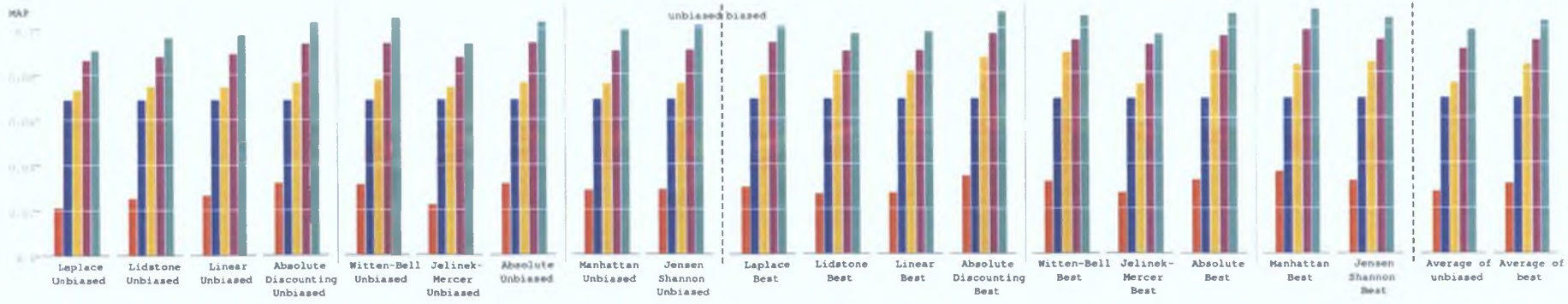
(c) TRECVid 2004 average results across retrieval models.

TRECVid 2004	*MAP*	*P@10*	*P@30*	*P@100*	*Impr.Text*	*Impr.Both*	*Impr.Oracle*
Text-Only	**.0686**	.209	.143	.091			
VisExs-CombMaxScore	**.0242**	.094	.058	.033			
Optimised	(**.0307**	.114	.070	.035)			
TextVis-CombWtScore	**.0752**	.219	.157	.100	+9.7%		
Optimised	(**.0833**	.258	.180	.096)	+21.4%		
TextVisBoth-Oracle	**.0903**	.279	.187	.106	+31.6%	+19.9%	
Optimised	(**.0942**	.303	.197	.107)	+37.3%	+13.3%	
TextVisComb-Oracle	**.0987**	.302	.196	.108	+43.9%	+31.1%	+9.3%
Optimised	(**.1026**	.316	.205	.110)	+49.6%	+23.4%	+8.9%

(a) TRECVid 2002



(b) TRECVid 2003



(c) TRECVid 2004

Figure 64: Results for combining the text and visual retrieval results using an *"TextVisBoth Oracle"* which chooses between text only, visual only or combined text and visual models for each topic and an *"TextVisComb Oracle"* which chooses the best mixing weights for combining text and visual results for each topic for retrieval models on the (a) TRECVid 2002, (b) TRECVid 2003 and (c) TRECVid 2004 collections. These Oracle-based results are compared with the visual-only *VisExs-CombMaxScore* results, the text-only results and the combined text and visual *TextVis-CombWtScore* results.

Table 27 Comparison of unbiased retrieval models for the TRECVid 2002, 2003 and 2004 search tasks for the *Vis-CombWtScore*, *VisExs-CombMaxScore* and *TextVis-CombMaxScore* fusion methods %Dif = difference from best result (lower better) R = rank of retrieval models (lower better) S = number of statistically significantly better retrieval models (lower better)

	Vis		*VisExs*		*TextVis*		*Average*	
Ret Method	*%Dif*	*R(S)*	*%Dif*	*R (S)*	*%Dif*	*R (S)*	*%Dif*	*R (S)*
Jensen-Shannon	-1 3%	2 0 (0 0)	-5 1%	2 7 (0 0)	-3 5%	5 3 (0 0)	**-3 3%**	**3 3(0 0)**
Manhattan	-3 8%	2 7 (0 7)	-6 6%	3 7 (0 3)	-3 5%	5 0 (0 0)	**-4 6%**	**3 8(0 3)**
Linear	-8 4%	5 7 (1 0)	-14 9%	4 0 (1 3)	-3 3%	6 3 (2 3)	**-8 9%**	**5 3(1 5)**
Lidstone	-6 1%	3 7 (0 0)	-18 3%	6 7 (2 0)	-2 8%	4 3 (1 0)	**-9 1%**	**4 9(1 0)**
Witten-Bell	-10 3%	5 0 (2 6)	-16 6%	4 7 (2 0)	-2 1%	3 7 (0 3)	**-9 7%**	**4 5(1 6)**
Absolute Discounting	-12 4%	6 7 (4 0)	-18 4%	6 0 (3 7)	-4 1%	6 0 (2 3)	**-11 6%**	**6 2(3 3)**
Laplace	-10 5%	4 7 (1 3)	-21 4%	5 3 (1 3)	-3 6%	6 0 (2 3)	**-11 8%**	**5 3(1 6)**
Jelinek-Mercer	-12 5%	6 7 (0 7)	-22 2%	5 7 (0 0)	-1 5%	3 3 (0 3)	**-12 1%**	**5 2(0 3)**
Absolute	-13 9%	8 0 (3 3)	-26 5%	6 3 (2 3)	-3 1%	5 0 (2 3)	**-14 5%**	**6 4(2 6)**

the fusion models This could be achieved by using query classes or by some automatic relevance feedback mechanism for updating the fusion weights

7 5 Discussion

We present a summary of all our synthesised results in Table 27 We can see clearly from this table that the visual results that use a single visual example (Vis columns) and the results that combine multiple visual examples (VisExs columns) are quite consistent with each other and that the differences between retrieval models are magnified when combining multiple visual examples We can also see that the text and visual fusion method (TextVis columns) achieves very similar performance for all retrieval models compared to the fusion of multiple visual examples The order of the results for the text and visual fusion is inconsistent with the order of results for the other two fusion tasks We believe that the Jensen-Shannon distance, Manhattan and the Jelinek-Mercer results are the best overall retrieval models though Jelinek-Mercer seems to perform relatively poorly in terms of MAP for the retrieval results when combining multiple visual examples

We believe our results are competitive with other visual-only approaches for the fully automatic TRECVid retrieval task The TRECVid 2004 workshop included fully automatic runs for the first time as a pilot scheme before hopefully their eventual more official inclusion in coming years Since only 4 groups DCU, Lowlands, CMU, IBM submitted runs in this pilot category, the results might not represent the true state-of-the-art in this research field Unfortunately, only two of these groups, DCU and Lowlands, submitted fully automatic visual-only results so it is even more difficult to gauge the state-of-the-art for visual-only performance on the TRECVid 2004 topics Recall that we have already compared our generative visual retrieval models with standard visual retrieval models such as Manhattan and Jensen-Shannon distance and that our language modelling approach performs slightly worse than Jensen-Shannon distance for visual-only retrieval, while for combined text and visual retrieval our language modelling approach

achieves better results than for these standards visual retrieval models

For the TRECVid 2004 search task Lowland's highest fully automatic visual-only run (LL-F-dyn-allvidm-RR, Gaussian Mixture Model of dynamic multi-spectral DCT texture) achieved a MAP of 0 010 (Ianeva et al , 2005), 55% lower than the submitted DCU visual language modelling run which achieved a MAP of 0 018 (actually 0 0175) by using a rank-based fusion of the Jelinek-Mercer colour, edge and texture results (Cooke et al , 2005) We also submitted a run based on Lidstone smoothing that achieved a MAP of 0 017, which was just slightly less than our submitted Jelinek-Mercer language model run

The unbiased results in this thesis improve dramatically on these visual results Our Jelinek-Mercer visual language model results, which is our second worst visual-only result on the TRECVid 2004 collection, achieves a MAP of 0 0222, a 27% improvement on our submitted Jelinek-Mercer rank-based fusion result and a 122% improvement on Lowland's submitted dynamic DCT GMM result Our best unbiased language modelling result for fully automatic visual retrieval on TRECVid 2004 is Absolute Discounting which achieves a MAP of 0 0319, which is a 82% improvement on our submitted Jelinek-Mercer results Our result for Lidstone smoothing is also improved achieving a MAP of 0 0250, a 47% improvement on the submitted Lidstone result Our full results for visual-only retrieval on the TRECVid 2004 collection are presented in Table 67 (appendix page 272) but are more easily compared as presented in Table 70 (appendix page 274) We *believe that our visual-only* results are at least competitive and most likely better than the continuous generative probability GMM approach

Four groups DCU, Lowlands, CMU and IBM submitted fully automatic combined text and visual runs for the TRECVid 2004 fully automatic search task The best result was submitted by DCU based on combining our rank-based Jelinek-Mercer visual result with the interpolated *shot+adj+video* language modelling text results, which achieved a MAP of 0 078 (actually 0 0776) The next best result was submitted by CMU and achieved a MAP of 0 075 The Lowlands multimodal result achieved an equally close MAP of 0 073 for combining their visual GMM results with their hierarchical text language model, which is the same type of text model that we use in our experiments These three combined text and visual runs were not statistically significantly different The IBM run achieved quite a lower MAP of 0 057 for the fusion of multimodal features (Amir et al , 2005) Our combined text and visual results are presented in Table 76 (appendix page 279) The score-based fusion of the Witten-Bell visual results achieves the same MAP of 0 0775 as our submitted run, however our score-based fusion of Jelinek-Mercer results achieves only a MAP of 0 0741, a 4 4% decrease on our submitted results The combination of the Lidstone visual language model with the text results produces a slightly higher MAP of 0 0745, which is 3 9% lower than our submitted results and is about the same as the submitted CMU result This decrease in our multimodal results occurs even though we have improved our visual-only results dramatically

Most of the difference between our submitted runs and our results in this thesis for multimodal search are due to the fact that in our submitted run we combined only the top 500 normalised visual results with the 1000 text results, whereas in this thesis we decided to treat all features the same and to always combine 1000 results from each retrieval model when performing text and visual fusion In future experiments we intend to look more closely at the combination of text and visual results with regard to the size of the result sets to be combined

In preliminary experiments we found the optimum number of visual results varied quite a lot between topics and collections and therefore we set it to a fixed 1000 results in this thesis in order to reduce noise that this parameter could produce We recognise that we have to look again at this decision and also investigate some other methods of multimodal result set fusion

7.6 Summary

We have shown that it is possible to successfully combine the visual features colour, edge and texture using weighted normalised scores The resulting combination is better than using any one of the features alone, such as colour, for the complete TRECVid retrieval task This fusion method is therefore a good default setting for visual retrieval for novice users or for initial query execution

We have shown that combining multiple visual examples is best achieved through the use of the maximum of normalised scores or normalised ranks (round-robin) The resulting combined results are better than the mean of the retrieval results for performing the query separately on each topic image but are significantly worse that a single image query that uses the best single visual example for each topic We need to perform further experiments on combining the top 2 and top 3 visual examples to verify if it is possible to improve on a single image query with the best visual example Also user experiments would be useful in order to understand whether users can select the best image reliably after a suitable exposure to the video collection

We have also shown that combining text and visual results, while resulting in an overall improvement in terms of MAP for TRECVid 2003 and TRECVid 2004, that this improvement is small at about 8% and 10% for the unbiased results for these two collections When we looked at the individual topic results we saw that in many cases the text results were superior but in a few the visual results were superior than the text alone

The visual retrieval models are useful for a few of the queries We believe that a query to a video collection should first be initiated with a text query if possible We have shown that for some queries the results can be improved by combination with visual models and for some others the visual query alone is best compared to the default TRECVid text queries for the given topics

We tested the visual results by using an Oracle that decided between combining them with the text results, using a visual-only query, or a text-only query We found that our visual results support up to 24 4% improvement for the TRECVid 2003 search task and up to 31 6% improvement for the TRECVid 2004 search task compared to only supporting text-only searching We believe these results support the inclusion of visual-based querying for access to video retrieval collections in that they show that visual querying can have a very positive effect on the retrieval performance

CHAPTER VIII

CONCLUSIONS

In this thesis we investigated the discrete language modelling approach for both text and visual retrieval of video shots, which we evaluated on the TRECVid 2002, 2003 and 2004 collections We believe our discrete visual language modelling approach represents a consistent, relatively effective and efficient approach to video shot retrieval In this chapter, we present our general conclusions for our discrete language modelling approach to video retrieval as well as more specific conclusions relating to each of the main aspects of our work – text-based, visual-based and combined multimodal retrieval of video shots We also look more closely at our evaluation methodology which provided the foundation of our experiments Research into discrete language models for video retrieval is, of course, unfinished and we suggest future work and extensions to our research

8 1 General Conclusions
8 2 Text Language Models
8 3 Visual Language Models
8 4 Fusion Methods
8 5 Evaluation Methodology
8 6 Future Work
8 7 Summary

8.1 General Conclusions

In this thesis we presented and evaluated the discrete language modelling approach for text and visual based retrieval of video shots This work extended the previous language modelling approaches for video retrieval (Jin and Hauptmann, 2002, Westerveld, de Vries and van Ballegooij, 2003) by considering different language modelling smoothing methods, by using discrete language models for both text and multiple visual features and by investigating different fusion approaches for combining the various text and visual feature language models

The benefit of our approach is that it is as effective as media-specific retrieval models, yet our approach uses a consistent retrieval model and representation for both text and visual features This work provides a foundation for multimedia retrieval that possibly can be more easily extended than other approaches as extensions to the discrete language modelling approach can be applied directly to both text and visual features Another benefit of our work is that we evaluated our discrete language modelling approach quite extensively in a wide range of experiments on three standard video search test collections These quite exhaustive fully-automatic empirical experiments provide a stronger basis, than is normally the case for video retrieval research, to form conclusions on the effectiveness of different retrieval models, feature representations and fusion methods

In this section we will highlight our general conclusions, while in the sections that follow we

will present more specific conclusions on the text language models, the visual language models, the fusion methods and our evaluation methodology. We will end this chapter with a description of future work and a brief final summary, but first our general conclusions:

Conclusion G.1 *The discrete text-based language modelling framework can be applied to visual retrieval for features such as colour, edges and texture.*

We applied the text-based language modelling approach to visual retrieval for the HSV colour, Canny edge and DCT texture features. We believe our approach is consistent for both mediums as the representation for ASR text and visual features are languages of discrete symbols and the retrieval models for both types of features are smoothed query-likelihoods. Our visual retrieval models can be further applied to many of the effective MPEG7 visual features (MPEG7 Scalable Colour, MPEG7 Colour Structure, MPEG7 Edge Histogram and MPEG7 GoF/GoP Colour descriptors) and to many popular visual features that are used in current video retrieval systems such as alternative colour features (e.g. Colour Coherence Vectors and Colour Correlograms), edge features (Edge Co-Occurrence Matrix and Edge Correlogram), texture features (Gabor response Histograms) and spatio-temporal features (Temporal Colour Correlogram, Temporal Gradient Correlogram).

Conclusion G.2 *The discrete language modelling approach is as effective as other retrieval models that are specific to the text or visual mediums.*

We compared language models and standard information retrieval models for the text-based retrieval of video shots and found that the language modelling approaches achieved in many cases better MAP than the TF-IDF and BM25 models, though these improvements were not statistically significant. We also compared language models and standard visual retrieval models for three visual features, HSV colour, Canny edge, and DCT texture, and found that the discrete language modelling approach achieves similar and sometimes better results than the standard visual retrieval models but the improvements with respect to the best standard visual retrieval model, Jensen-Shannon distance (Jeffrey Divergence), were again not statistically significant. We believe that our empirical evidence confirms that discrete language models are as good as standard retrieval models for either medium.

Conclusion G.3 *The discrete visual language modelling approach is as efficient as other standard visual retrieval models.*

Our discrete language modelling approach is as efficient as current approaches that use standard visual retrieval models with multidimensional histograms. The time complexity of the discrete query-likelihood retrieval models, except for the Bayesian language model, is related to the number of distinct symbols in the query's visual example (symbols or histogram bins with non-zero counts), whereas for standard distance measures such as Jensen-Shannon distance (Jeffrey Divergence) the execution speed is related to the size of the query language (number of cells in the multidimensional histogram) and the speed of the continuous language models such as the DCT GMM approach is related to the number of query samples. The number of distinct query symbols is upper-bounded by the size of the visual language and the number

of query samples and therefore the discrete query-likelihood method compares very favourably with these approaches Furthermore, in a static video collection or a dynamic collection where it is feasible to completely rebuild a snapshot of the index, the smoothed log-probabilities for the visual features can be stored in the index, leading to more efficient calculation of the query log-likelihood

Conclusion G 4 *The unsmoothed MLE probability model is ineffective for both text-based and visual-based retrieval*

This is an expected outcome since the MLE model will score all documents that are missing a single query term with zero probability We have verified this in our experiments for text and most visual features The only exceptions are visual features with very small visual languages such as the tested global Canny features, the largest of which is *Canny 64+1* that has only 65 symbols, which since they have very few zero frequency terms in their document representations do not produce better results with smoothed query-likelihoods

Conclusion G 5 *The combination-based (interpolation-based) language models perform better than the discounting-based language models for both text and most regional visual features*

The difference between discounting and combination-based smoothing is less apparent for regional visual features than for text, while for global visual features there is no consistent difference The combination-based smoothing methods perform better for the regional Canny edge and regional DCT texture features, while the discounting-based smoothing methods work slightly better for the regional HSV colour feature This difference may be due to Canny and DCT features having more skewed collection distributions than the HSV colour feature and therefore are more likely to benefit from smoothing with the background distribution The difference between the Jelinek-Mercer combination-based language model and the best discounting models Laplace, Lidstone and Linear is only 2% (0 0087 compared to 0 0089) for the regional colour feature on the TRECVid 2004 collection, which could hardly be considered significant and the larger differences on the TRECVid 2002 and TRECVid 2003 collections are attributable to the very different visual colour quality on these two collections, which leads to inappropriate unbiased smoothing parameters Overall, all visual features perform quite poorly compared to text and the absolute difference between the different visual retrieval models is very small in terms of MAP

Conclusion G 6 *The fusion of text and visual results over all topics achieves negligible improvement for the TRECVid 2002 collection and on average achieves only 8% and 10% better MAP compared to text-only results for the TRECVid 2003 and TRECVid 2004 collections*

These somewhat negative fully-automatic retrieval results under-represent the performance of visual search in a video retrieval system because the improvement in some topic's results is counter-balanced by the decrease in the performance of other topics for which text-only searching is the best strategy The fully-automatic fusion results are as much a measure of the negative effects of the visual results on the good text-oriented topics than the positive effects it has on other topics This problem is due to our fusion methods treating all topics identically and an

automatic adaptive method of combining text and visual feature results, such as Co-Retrieval (Yan and Hauptmann, 2004), may improve our fully-automatic text and visual fusion results

Conclusion G 7 *The ideal choice between text-only, visual-only and combined text and visual search for each of the default TRECVid topics could support a potential increase in terms of MAP of 6 0% on TRECVid 2002, 24 4% on TRECVid 2003 and up to 31 6% on the TRECVid 2004 collection compared to text-only search results*

These results indicate a large improvement in the TRECVid 2003 and TRECVid 2004 results by an Oracle-based selection between text-only, visual-only or combined text and visual search strategies We believe these results support the inclusion of visual-based querying for access to video collections as our results show that visual querying has a very positive effect on retrieval performance when selectively combined with text retrieval methods User-based experiments are required in order to more appropriately support this claim

8.2 Text Language Models

In this thesis our contribution to the research into text-based approaches to video retrieval is to propose extensions to the hierarchical language modelling approach (Westerveld, de Vries and van Ballegooij, 2003) that use different smoothing models and to evaluate different structural representations based on the physical and semantic video structures for the video shot retrieval task We also establish a baseline for standard non-hierarchical language models and standard information retrieval models for the text-based video shot retrieval task

Conclusion T 1 *The* adj-only *text representation, a window of adjacent shot's text, is the best of the tested physical non-hierarchical representations for text-based video shot retrieval*

Conclusion T 2 *The semantic* story-only *text representation is the best of the tested non-hierarchical representations for text-based video shot retrieval*

The *story-only* representation provides statistically significantly better results than *adj-only*, but similar to *adj-only* its results are not statistically significantly better than the *shot-only* representation This indicates that while the *story-only* and *adj-only* text representations improve the average performance compared to the *shot-only* representation for the video retrieval tasks, it does so unreliably across the TRECVid topics

Conclusion T 3 *The* shot+adj+video *representation is the best tested hierarchical physical representation for text-based video shot retrieval*

The *shot+adj+video* representation achieves statistically significantly better results than the *shot-only* and *adj-only* representations and it also improves on the *story-only* representation, though this improvement is not statistically significant

Conclusion T 4 *The semantic* shot+story *hierarchical representation is the best tested representation for text-based retrieval of video news shots*

The *shot+story* representation is statistically significantly better than the *shot-only*, *adj-only* and *shot+adj+video* representations The 10% improvement for the *shot+story* representation relative to the best physical representation *shot+adj+video* is less likely to be achieved when applying a semantic hierarchical structure to other types of video content since news programmes are a special case where each semantic story unit is usually quite distinct from others in the same programme

Conclusion T 5 *The combination-based language models are better than the discounting-based language models for text-based retrieval using either the best non-hierarchical physical representation* adj-only *or the even better semantic* story-only *representation*

This outcome is in line with other results for text-based information retrieval systems Unlike visual-based features, for text-based features there is a notable difference between discounting-based language models and combination-based language models that smooth using the background collection model

Conclusion T 6 *The Dirichlet, Bayesian and Absolute language models achieve similar and sometimes slightly better results than the Jelinek-Mercer language model for the non-hierarchical text-based representations*

The Dirichlet, Bayesian and Absolute language models achieve better results than the Jelinek-Mercer language model for the *adj-only* representation and for some video search collections these differences are statistically significant We found that the Jelinek-Mercer language model produces the best results for the *story-only* representation, though it was not statistically significantly better than the other combination-based language models We also found that the Dirichlet and Bayesian language models achieve slightly higher MAP than BM25 for the *adj-only* representation, though again these differences were never statistically significant for any of the tested collections We view these results as an indication that language models other than the Jelinek-Mercer language model have a potential benefit in the video shot retrieval task

Conclusion T 7 *Our proposed hierarchical language models, hierarchical Absolute, hierarchical Witten-Bell, and hierarchical Dirichlet, do not perform as well as the hierarchical Jelinek-Mercer (linear interpolation) language model*

Our proposed hierarchical language models perform worse than the hierarchical Jelinek-Mercer language model (Westerveld, de Vries and van Ballegooij, 2003) for the best hierarchical physical representation *shot+adj+video* and the best hierarchical semantic representation *shot+story* The results for the semantic hierarchical representations are a lot closer than the physical hierarchical representations and indicate a potential equivalence in performance between the hierarchical Jelinek-Mercer, hierarchical Witten-Bell and hierarchical Absolute language models, while the hierarchical Dirichlet language model in general performs statistically significantly worse than hierarchical Jelinek-Mercer for the tested hierarchical representations Since

the *adj* and *video* representations introduce topical noise into the hierarchical *shot+adj+video* representation for the two news collections, TRECVid 2003 and TRECVid 2004, keeping their influence constant between different document representations, as in the hierarchical Jelinek-Mercer language model, seems to be a beneficial strategy

8.3 Visual Language Models

In this thesis we applied the discrete language modelling approach to the visual HSV colour, Canny edge and DCT texture features We evaluated different discrete language models and different global and regional representations for these three visual features on the TRECVid 2002, 2003 and 2004 collections

Conclusion V 1 *The best visual feature for the video search task is regional colour, then regional texture and finally regional edge*

This is an average result across retrieval models for the set of topic visual examples on the three video collections and for individual visual examples different features are best Indeed for the recent TRECVid 2004 search task the differences between the colour and texture feature is less acute and is actually reversed though by only a small amount in terms MAP for some retrieval models

Conclusion V 2 *The MPEG7 inspired* HSV 16x4x4 *global colour representation is the best of the tested HSV colour representations*

Conclusion V 3 *The Canny 64+1 edge features is the best of the tested global Canny representations*

The *Canny 32+1* representation achieves similar results to the *Canny 64+1* representation and a higher number of edge orientations may produce slightly better results but we have not tested this This result was unexpected and highlights the lack of discrimination power in the small global Canny representation of 4 edge orientations The performance differences between Canny representations is likely diminished when considering regional variations and in fact lower order Canny features with less that 64 edges orientations may be preferable in those configurations

Conclusion V 4 *The TRECVid 2002 and TRECVid 2003 collections are visually very different and this caused inconclusive results between these collections as to which of the tested DCT texture representations was best*

For TRECVid 2002 the best representation was *DCT 4x4x4x4* whereas for TRECVid 2003 the best representation was *DCT 8x8x8* We chose the *DCT 3x3x3x3x3* representation for our regional experiments due to its relative performance on the TRECVid 2003 collection and due

to the fact that it models more spatial frequency than the other representations that were better than it on the TRECVid 2002 collection.

Conclusion V.5 *The 5x5 regional feature representations achieves generally better results than the global and other tested regional representations for the three visual features.*

This is true for all the features on the three collections except the colour and edge features on TRECVid 2002 collection. On this collection global colour is better than regional colour representations and the regional 3x3 edge representation is better than the 5x5 regional edge representation. For the other search tasks more regions such as 6x6 or even 10x10 may produce better results, however we have not yet evaluated this and the execution speed would be prohibitive, four times slower for the 10x10 regions than for the 5x5 regions, unless we reduce the size of the representation of each of the colour, edge and texture features. A non-uniform partitioning of the x and y dimension so that there is more partitions in one dimension such as the x dimension may also be beneficial.

Conclusion V.6 *For the smoothed language models, the discounting-based language models work slightly better for regional HSV colour features, while the combination-based language models work slightly better for the regional Canny edge and regional DCT texture features.*

The difference between the different language model's results is a lot smaller for the visual features than for the text features and unlike text the results for visual features do not have a consistent separation in performance terms between discounting-based language models and the interpolation-based language models. Both Canny edge and DCT texture feature have languages that are less uniform than the HSV colour feature and this could explain why smoothing with the background collection model is more important for these features.

Conclusion V.7 *Jensen-Shannon distance (Jeffrey Divergence) is the best of the tested standard visual retrieval models for most visual features.*

In terms of traditional visual retrieval models Jensen-Shannon distance (Jeffrey Divergence) is the best, Manhattan distance is in second place and Euclidean distance is consistently the worst. Jensen-Shannon distance has similarities to the language modelling approach but does not require external smoothing as it models the hypothetical common source using a midpoint distribution of the query and document. Unlike query-likelihood or document-likelihood, it is a symmetric distance measure that quantifies how well both the query and document match this common source. In essence it combines query and document-likelihood but in terms of the relative entropy measure. Though not used in most video retrieval systems, which actually predominantly use Euclidean or Manhattan distance, we believe that we have shown that Jensen-Shannon distance is highly suited to the task of feature matching for general video search. Euclidean distance performs worse than the other matching models primarily due to the fact that it magnifies the difference between visual symbols (histogram bins) by squaring the difference when comparing two images. This may be appropriate for exact match searching but for general video retrieval at least for our tested features it is not desirable. The MLE model is actually the only retrieval model to consistently produce worse results than Euclidean distance on the

individual features Lower-order Minkowski distances such as Fractional distances (Howarth and Ruger, 2005) may close the performance gap with Jensen-Shannon distance, but we have as yet not evaluated these measures

Conclusion V 8 *The Jelinek-Mercer language model is the best language model for separate visual feature search and achieves results that are overall as good as the best standard visual retrieval model*

The Jelinek-Mercer language model produces the best results for the *Canny 5x5 64+1* regional feature and the Jensen-Shannon distance (Jeffrey Divergence) produces on average 15% worse results for this feature on the three TRECVid collections The Jelinek-Mercer language model achieves the third best overall result for the *DCT 5x5 3x3x3x3x3* feature, but is in fact best for both the TRECVid 2002 and TRECVid 2003 collections and is on average 9% better than Jensen-Shannon distance The Jelinek-Mercer language model produces slightly poor results for the colour feature achieving 10 6% poorer MAP than the Jensen-Shannon distance On the regional HSV colour feature the discounting-based language models actually perform better than the combination-based language models and Laplace and Lidstone achieve on average only 2% less MAP than Jensen-Shannon distance Since all our document representations have the same number of samples (keyframes have the same dimensions), both Laplace and Lidstone can be thought of as interpolation with a uniform source unlike the Jelinek-Mercer language model which is an interpolation with a background collection We believe that Jelinek-Mercer smoothing is the overall the best choice if picking the same smoothing model for all features

Conclusion V 9 *The structural smoothing methods that combine the shot representation with adjacent shots, stories or the video structural units are not beneficial to visual-based video retrieval for the TRECVid 2003 news content or for the majority of topics on the TRECVid 2002 collection*

The results when using video structure are extremely disappointing for visual features since the visual features represent nothing distinct when built from the adjacent shot's keyframes or the whole video's keyframes This is perhaps not surprising when we consider the high degree of visual dissimilarity between adjacent shots

8 4 Fusion Methods

In this thesis we investigated different fusion methods for combining our language model results for text and visual features The standard fusion methods consist of combining the results based on independent probability and their normalised scores or normalised ranks using either the average, weighted average, or maximum functions The probability-based fusion method simply assumes independence between features and visual examples and therefore multiplies their probabilities (or equivalently adds their log-likelihoods)

Conclusion F 1 *The results for multiple features are best combined using a weighted average of normalised scores*

This indicates that scores (the normalised top 1000 log-probabilities) can be fused together in a positive way where each feature plays a cooperative role in scoring documents in the fusion's result list In follow on experiments, we identified instability in the weighting of different features between TRECVid collections and found that combining using CombSUM (summing normalised scores without weights) achieves overall better results (Mc Donald and Smeaton, 2005)

Conclusion F 2 *The results for multiple visual examples are best combined using the maximum of normalised scores or ranks*

This implies that the results from the different visual examples that have previously been fused for multiple features cannot be reliably combined in a way where the scores for the different visual examples refine and corroborate each other Combining based on maximum normalised score produced marginally better results than maximum normalised rank (round-robin) for combining multiple visual examples The closeness to the round-robin results suggests that each visual example represents a separate visual impression of the relevant items which cannot be more directly combined with each other

Conclusion F 3 *The text and visual results are best combined using the weighted average of normalised scores*

The weighted normalised score fusion method produces marginally better results than weighted normalised rank for combining text and visual results The similarity between normalised rank and normalised scores is not surprising since the visual scores are already quite linear and resemble normalised ranks The linear nature of the visual scores is caused by the previous application of the normalised score fusion methods to combine visual features and visual examples In comparison combining text and visual scores using joint probability results in as poor results as the visual model alone, because its generative probabilities are many orders of magnitude more dominant in the combination than for the small sample of query text

Conclusion F 4 *The visual-only results are improved by between 20% to 40% on the three video search test collections if we represent each topic using the single best image per topic instead of fusing all the visual example searches*

It is not apparent to us whether a user, who does not have an intimate knowledge of the collection being searched, would be able to select the single best image for each topic Therefore we believe fusing the multiple results from the different visual examples is a good initial strategy Relevance feedback could use the knowledge of some examples being better than others to improve the ranking Combining visual examples using maximum normalised score or maximum normalised rank is a quite negative (almost non-fusion) approach to combining these results sets and this indicates why the single best image may produce a better ranking than the fusion of results from a set of visual examples

8.5 Evaluation Methodology

In this thesis we compared retrieval models and feature representations using fully automatic experiments for three standard TRECVid test collections Our evaluation procedure involved tuning all parametric retrieval models in an unbiased manner by optimising their parameters in terms of the MAP measure on a separate collection

Conclusion E 1 *The TRECVid collections and multimedia topics support the fully automatic evaluation of different retrieval models and feature representations in controlled and repeatable experiments*

Unlike the official TRECVid manual experiments, fully automatic experiments as provisionally supported by the TRECVid 2004 initiative provide a means through which different researchers can directly compare results The current TRECVid manual and interactive experiments are far more difficult to interpret or compare between retrieval groups The investigation of retrieval models and alternative feature representations, as in this thesis, is ideally suited to fully automatic and repeatable experiments Of course interfaces, more advanced query formulations or different user scenarios are either best suited to user-based interactive or manual experiments

Conclusion E 2 *The tuning of parameters for visual-based search by optimising the MAP measure can result in over-fitting the parameters on a minority of topics, which produces unsuitable parameters for the unbiased testing on another collection*

The performance of visual-only searching on the TRECVid collections is highly variable and a small minority of one or two topics can dominate the MAP statistic, thereby skewing the parameter selection totally to just these topics This can lead to either underestimating the amount of smoothing or choosing a totally inappropriate smoothing parameter value The inappropriate parameter settings are often caused by topics with very few relevant results, which can achieve a very high average precision from the inclusion of any of its relevant results in the top ranked documents This problem with optimising using the MAP measure is present in other information retrieval tasks when the set of topics are very difficult or vary significantly in difficultly level Machine learning optimisation methods and other solutions to tuning parameters such as dampening each topics average precision with a suitable monotonic function before calculating MAP may be more effective than using MAP directly It should be kept in mind that the optimisation criteria and the evaluation criteria need not be the same

Conclusion E 3 *Performance measures, such as MAP or overall precision at N documents, that average over the TRECVid topics can be unreliable measures for comparing the retrieval results of visual-only searching*

Since the visual-only retrieval results tend to be very poor with a few topics dominating the unbiased results, these measures can sometimes represent the performance of the best one or two topics more than the performance across the whole set of topics This problem with using the mean performance in visual-only results implies that we need to use statistical tests or to

compare results individually on topics in order to identify reliable improvements Statistical tests are a useful mechanism to sort through the differences between average performance measures in order to identify the more meaningful or more consistent differences between retrieval results from those differences that are likely due to chance In some cases looking at the individual topic results can be informative as to the differences between retrieval runs This observation implies that researchers will find it difficult to properly compare their results with published results unless they have access to individual topic results in order to use statistical tests or to compare them individually on each topic The problem of interpreting mean performance of visual-only runs over the set of TRECVid topics can be alleviated by using multiple TRECVid collection when reporting experiments Furthermore the results of multiple search tasks may be aggregated together to achieve more reliable statistics on which both to compare mean performance and to perform statistical tests Ideally each years TRECVid experiment would contain more than 25 topics to make the visual-only results more reliable but aggregating results from subsequent years provides a pragmatic solution

Conclusion E 4 *The unbiased fully automatic evaluation strategy suppresses the measurable benefits of the fusion of visual-based searching with text-based searching*

The combination of text searching with visual searching is a beneficial strategy for some topics whereas for other topics it decreases retrieval performance compared to text searching alone Some very good text-based topics can be adversely effected decreasing the MAP (or other measures) by such an amount that the improvement in the other topics is largely swallowed up Text-only searching and combined text and visual searching are strategies to find video shots and we believe that combined text and visual searching need not be consistently better than text-only searching for every topic in order for users to consider it useful We found that Oracle-based runs, where an Oracle makes the best choice between the different search strategies, such as between text-only, visual-only and combined text and visual searching, provide a useful fully automatic experiment showing the potential benefit of visual search in the video shot retrieval task However, these Oracle-based runs have limited use as they are biased and therefore provide unreliable results for the direct comparison of retrieval models

8.6 Future Work

The discrete language modelling approach presented in this thesis can be extended to support updating and reweighting of the text and visual query representations As described in chapter 3 we can use relative entropy (or equivalently KL divergence), which supports the updating of the query representation within a relevance feedback framework, to score document language models relative to a query language model (Zhai and Lafferty, 2001a) Of course since both text and visual features use the same discrete language model representation, any relevance feedback approach for the text feature can also be applied to our visual features Related to the notion of relevance feedback is the Co-Retrieval approach of (Yan and Hauptmann, 2004), which automatically incorporates additional features based on their consistency with the initial query results This Co-Retrieval approach was successfully applied to the TRECVid 2003 search topics using the Euclidean distance measure and it would be interesting to investigate it for the combination of our generative probabilities of each of the visual features A simpler approach

to an automatic reweighting mechanisms is to classify search requests into some broad ontology (eg finding people, objects, natural scenes, etc (Hauptmann et al , 2004)) and to tune our retrieval approaches to these specific types of queries This would allow all default weighting of the different features to be more appropriate for the type of query and may provide a more solid basis for automatic relevance feedback or feature boosting

We believe that our text experiments in chapter 5 should be performed on other text descriptions of video shot such as closed captions, screen captions, and other recognised video text This would provide a wider understanding of the benefits of text-based searching in video retrieval We are also currently investigating extensions to our proposed hierarchical language models to consider hybrid hierarchical smoothing where the individual levels in the hierarchy use their own smoothing models and are combined with other levels in the hierarchy using a linear interpolation This hybrid approach may better support alternative smoothing methods for hierarchical video representations than the proposed hierarchical language models in this thesis It would also be informative to benchmark the standard retrieval models BM25 and TF-IDF for the hierarchical structures so as to support comparison with the hierarchical language models

Our experiments into visual languages for HSV colour, Canny edge and DCT texture histogram features in chapter 6 should be further extended to other types of visual feature representations such as co-occurrence matrices, coherence vectors and correlograms Gabor wavelets are a more popular and successful texture representation than DCT coefficients and it would informative to see whether histogram representation of Gabor wavelet responses can be successfully used in the discrete language modelling approach to visual retrieval

We believe that a comprehensive study of the MPEG7 visual features for general video search would be a worthwhile contribution to the video retrieval field This standard provides compact visual features such as MPEG7 Scalable Colour, MPEG7 Colour Structure, MPEG7 Edge Histogram and MPEG7 GoF/GoP Colour descriptors that are amenable to the language modelling approach Retrieval models for other features m the standard and fusion methods should be investigated with the TRECVid video retrieval task, which was unavailable during the development of this standard

It would also be interesting to investigate language models for extreme representations such as with a large number of regions, or for very large non-regional visual languages to tease out the role of smoothing in sparser visual representations Unlike text language models, visual features have dimensions with ordinal scales that may provide an extra avenue for smoothing visual terms based on the local neighbourhood of similar valued features A limited version of this smoothing, called local-smoothing, was applied in the original discrete generative colour model for video retrieval (Jin and Hauptmann, 2002), which smoothed a region's colour model with colour models from spatially close regions

A further consideration in visual search is non-keyframe based video shot representations and our approach which simply represents each shot with a keyframe's visual features can be applied to indexing key-segments, moving regions or objects within a *shot* Our preliminary investigations of spatio-temporal descriptions did not show much improvement compared to the keyframe based approach We think structural smoothing may be useful at the sub-shot level to combine the probability models of the moving regions (or objects) with the probability models

for the whole shot's visual contents with a somewhat smaller enclosing temporal neighbourhood (i e with a fixed number of adjacent frames)

Our research into fusion methods for video search in chapter 7 could be further expanded to establish whether the fusion of the top 2 or 3 visual examples achieves better results than searching using the single best visual example for each topic We could also investigate whether combining different result set sizes for visual features improves performance consistently across topics and collections Furthermore, it would be interesting to look into score distribution modelling (Manmatha et al , 2001) and regression methods for combining the different results

Our optimisation strategy for the retrieval model's parameters requires further consideration We believe we need to identify a better optimisation criteria for our retrieval models since our current strategy of optimising MAP leads to over-fitting of some of the language models on a minority of visual topics, which we believe may have contributed to poorer unbiased results We intend to look at normalising the individual topic's results in order to reduce the effect of one or two topics dominating the selection of the smoothing parameters Since the TRECVid 2003 and TRECVid 2004 collections are so similar it would be interesting to cross-validate the parametric language models on both collections, which may result in more consistent and clearer unbiased results for TRECVid 2003 since we currently choose its parameters based on the very different TRECVid 2002 collection So far our evaluation strategy is system-oriented and it would be prudent to compare search strategies such as automatic fusion of features versus choosing the importance of each feature, or the choice between text, visual, or both text and visual retrieval compared with manually setting the importance of text and visual features in user-based experiments We believe that our system-oriented experiments provide a firm foundation on which to build upon in user-based experiments in order to get a clearer understanding of what works or doesn't work for users involved in the video retrieval task

8.7 Summary

In this thesis we extended the range of discrete language modelling based information retrieval from text to full multimedia retrieval and showed through exhaustive empirical experiments on three video test collections that this consistent approach emulates the retrieval performance achieved by media-specific retrieval models The importance of this research is that extensions to the discrete language modelling approach can be applied to both text and visual based video retrieval, thereby doubling the scope of the benefits This research also provides a complementary viewpoint on the traditional histogram-based retrieval approaches to video retrieval

This thesis only touches at some of the potential for applying this consistent discrete language modelling retrieval approach to both text and visual based retrieval of video content The discrete language modelling approaches presented in this thesis can be further extended, for example through the use of relevance feedback, ordinal-based smoothing of visual features, object segmentation and better probabilistic modelling for the combination of the different feature's language modelling results

APPENDIX A

TRECVID TOPICS

Table 28 TRECVid 2002 search topics I = Number of images in topic description V = Number of Videos in topic description Rel = Number of relevant shots for topic in collection Classification is reproduced from (Smeaton and Over, 2003) and is based on a classifications of visual topics (Armitage and Enser, 1996) into general and specific requests for persons or things (PT), events (E) or places (P)

					Specific		*Generic*		
T	**Text Description**	**I**	**V**	**Rel**	PT	P	PT	E	P
75	Find shots with Eddie Rickenbacker in them	2	2	15	PT				
76	Find additional shots with James H Chandler	0	3	47	PT				
77	Find pictures of George Washington	1	1	3	PT				
78	Find shots with a depiction of Abraham Lincoln	1	1	6	PT				
79	Find shots of people spending leisure time at the beach, for example walking, swimming, sunning, playing in the sand Some part of the beach or buildings on it should be visible	0	4	55			PT	E	P
80	Find shots of one or more musicians a man or woman playing a music instrument with instrumental music audible Musician(s) and instrument(s) must be at least partly visible sometime during the shot	0	2	63			PT	E	
81	Find shots of football players	0	4	15			PT		
82	Find shots of one or more women standing in long dresses Dress should be one piece and extend below knees The entire dress from top to end of dress below knees should be visible at some point	0	3	170			PT	E	
83	Find shots of the Golden Gate Bridge	5	0	33	PT	P			
84	Find shots of Price Tower, designed by Frank Lloyd Wright and built in Bartlesville, Oklahoma,	1	0	4		P			
85	Find shots containing Washington Square Park's arch in New York City The entire arch should be visible at some point	0	1	7		P			
86	Find overhead views of cities - downtown and suburbs The viewpoint should be higher than the highest building visible	0	4	105			PT		P
87	Find shots of oil fields, rigs, derricks, oil drilling/pumping equipment Shots just of refineries are not desired	0	1	40			PT		P
88	Find shots with a map (sketch or graphic) of the continental US	0	4	72			PT		
89	Find shots of a living butterfly	2	0	10			PT	E	
90	Find more shots with one or more snow-covered mountain peaks or ridges Some sky must be visible them behind	0	3	75			PT		P
91	Find shots with one or more parrots	1	1	17			PT		
92	Find shots with one or more sailboats, sailing ships, clipper ships, or tall ships - with some sail(s) unfurled	3	2	47			PT		
93	Find shots about live beef or dairy cattle, individual cows or bulls, herds of cattle	0	5	161			PT		
94	Find more shots of one or more groups of people, a crowd, walking in an urban environment (for example with streets, traffic, and/or buildings)	0	3	303			PT	E	P
95	Find shots of a nuclear explosion with a mushroom cloud	0	3	17			PT	E	
96	Find additional shots with one or more US flags flapping	0	2	31	PT		PT		
97	Find more shots with microscopic views of living cells	0	2	82			PT	E	
98	Find shots with a locomotive (and attached railroad cars if any) approaching the viewer	0	5	56			PT	E	
99	Find shots of a rocket or missile taking off Simulations are acceptable	0	2	11			PT	E	

Table 29 TRECVid 2003 search topics I = Number of images in topic description V = Number of Videos in topic description Rel = Number of relevant shots for topic in collection Classification is reproduced from (Kraaij et al , 2004) and is based on a classifications of visual topics (Armitage and Enser, 1996) into general and specific requests for persons or things (PT), events (E) or places (P)

					Specific		*Generic*		
Top	**Text Description**	**I**	**V**	**Rel**	PT	P	PT	E	P
100	Find shots with aerial views containing both one or more buildings and one or more roads	4	4	87			PT		
101	Find shots of a basket being made - the basketball passes down through the hoop and net	2	4	104			PT	E	
102	Find shots from behind the pitcher in a baseball game as he throws a ball that the batter swings at	0	5	183			PT	E	
103	Find shots of Yasser Arafat	1	0	33	PT				
104	Find shots of an airplane taking off	1	2	44			PT	E	
105	Find shots of a helicopter in flight or on the ground	4	2	52			PT	E	
106	Find shots of the Tomb of the Unknown Soldier at Arlington National Cemetery	4	0	31	PT	P			
107	Find shots of a rocket or missile taking off Simulations are acceptable	4	4	62			PT	P	
108	Find shots of the Mercedes logo (star)	3	0	34	PT				
109	Find shots of one or more tanks	2	2	16			PT		
110	Find shots of a person diving into some water	3	1	13			PT	E	
111	Find shots with a locomotive (and attached railroad cars if any) approaching the viewer	3	4	13			PT	E	
112	Find shots showing flames	3	4	228			PT		
113	Find more shots with one or more snow-covered mountain peaks or ridges Some sky must be visible them behind them	3	2	62			PT		P
114	Find shots of Osama Bin Laden	3	0	26	PT				
115	Find shots of one or more roads with lots of vehicles	5	4	106			PT		P
116	Find shots of the Sphinx	3	0	12	PT				
117	Find shots of one or more groups of people, a crowd, walking in an urban environment (for example with streets, traffic, and/or buildings)	4	4	665			PT	E	P
118	Find shots of Congressman Mark Souder	2	0	6	PT				
119	Find shots of Morgan Freeman	3	0	18	PT				
120	Find shots of a graphic of Dow Jones Industrial Average showing a rise for one day The number of points risen that day must be visible	0	6	47	PT				
121	Find shots of a mug or cup of coffee	3	2	95			PT		
122	Find shots of one or more cats At least part of both ears, both eyes, and the mouth must be visible The body can be in any position	4	3	122			PT		
123	Find shots of Pope John Paul II	5	2	45	PT				
124	Find shots of the front of the White House in the daytime with the fountain running	2	3	10			PT	E	P

Table 30 TRECVid 2004 search topics I = Number of images in topic description V = Number of Videos in topic description Rel = Number of relevant shots for topic in collection Classification is reproduced from (Kraaij et al , 2004) and is based on a classifications of visual topics (Armitage and Enser, 1996) into general (Gen) and specific requests for persons or things (PT), events (E) or places (P)

					Specific	*Generic*		
Topic	**Text Description**	**I**	**V**	**Rel**	PT	PT	E	P
125	Find shots of a street scene with multiple pedestrians in motion and multiple vehicles in motion somewhere in the shot	1	2	154		PT	E	P
126	Find shots of one or more buildings with flood waters around it/them	2	4	118		PT		
127	Find shots of one or more people and one or more dogs walking together	0	6	64		PT	E	
128	Find shots of U S Congressman Henry Hyde's face, whole or part, from any angle	5	1	115	PT			
129	Find shots *zooming* in on the U S Capitol dome	2	3	16	PT			
130	Find shots of a hockey rink with at least one of the nets fully visible from some point of view	2	3	162		PT		P
131	Find shots of fingers striking the keys on a keyboard which is at least partially visible	0	4	86		PT	E	
132	Find shots of people moving a stretcher	0	5	41		PT	E	
133	Find shots of Saddam Hussein	3	2	46	PT			
134	Find shots of Boris Yeltsin	3	4	22	PT			
135	Find shots of Sam Donaldson's face - whole or part, from any angle, but including both eyes No other people visible with him	1	4	54	PT			
136	Find shots of a person hitting a golf ball that then goes into the hole	0	3	19		PT	E	
137	Find shots of Benjamin Netanyahu	4	4	106	PT			
138	Find shots of one or people going up or down some visible steps or stairs	4	4	97		PT	E	
139	Find shots of a handheld weapon firing	4	4	55		PT	E	
140	Find shots of one or more bicycles rolling along	3	3	69		PT	E	
141	Find shots of one or more umbrellas	5	5	54		PT		
142	Find more shots of a tennis player contacting the ball with his or her tennis racket	3	3	41		PT	E	
143	Find shots of one or more wheelchairs They may be motorized or not	4	4	39		PT	E	
144	Find shots of Bill Clinton speaking with at least part of a U S flag visible behind him	2	2	96	PT	PT	E	
145	Find shots of one or more horses in motion	2	5	67		PT	E	P
147	Find shots of one or more buildings on fire, with flames and smoke visible	0	4	85		PT	E	
148	Find shots of one or more signs or banners carried by people at a march or protest	5	6	194		PT		

APPENDIX B

ADDITIONAL TABLES FOR CHAPTER 5

Table 31: Comparison of the *adj-only* ASR text representation with the *shot-only* and *video-only* representations for the different retrievals on the TRECVid 2002 collection.

TRECVID 2002		*Adj-Only*				*V. Shot-Only*	*V. Video-Only*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*
MLE (Adj)	4	**.0262**	.060	.013	.000	+29.3% ∼ .233	+31.3% ∼ .250
Laplace (Adj)	4	**.0978**	.180	.140	.089	+53.4% ∼ .012	+61.8% ∼ .079
Natural (Adj)	4	**.0859**	.140	.135	.086	+33.1% ∼ .014	+23.9% ∼ .274
Lidstone (Adj, λ)	4, 0.06	**.1022**	.176	.151	.088	+58.8% ∼ .005	+66.6% ∼ .066
Best	(2, 0.34	**.1287**	.216	.168	.079)	+96.5% ∼ .001	+83.1% ∼ .112
Linear (Adj, α)	4, 0.07	**.0995**	.176	.128	.088	+55.8% ∼ .013	+43.3% ∼ .177
Best	(1, 0.04	**.1160**	.224	.164	.071)	+68.3% ∼ .000	+57.3% ∼ .307
Absolute Discounting (Adj, δ)	4, 0.83	**.1069**	.180	.148	.089	+73.1% ∼ .003	+54.9% ∼ .125
Best	(2, 0.89	**.1256**	.220	.163	.079)	+85.8% ∼ .001	+77.6% ∼ .106
Witten-Bell (Adj)	4	**.1184**	.200	.135	.092	+70.4% ∼ .010	+68.5% ∼ .100
Jelinek-Mercer (Adj, λ)	4, 0.02	**.1069**	.204	.137	.092	+53.8% ∼ .011	+55.3% ∼ .084
Best	(1, 0.67	**.1220**	.224	.159	.075)	+73.4% ∼ .000	+73.5% ∼ .204
Absolute (Adj, δ)	4, 0.85	**.1221**	.200	.152	.093	+77.5% ∼ .004	+74.8% ∼ .054
Best	(2, 0.84	**.1290**	.236	.165	.083)	+83.3% ∼ .000	+80.3% ∼ .094
Dirichlet (Adj, μ)	4, 750	**.1166**	.184	.140	.086	+69.2% ∼ .007	+80.4% ∼ .047
Best	(1, 100	**.1268**	.236	.159	.075)	+82.0% ∼ .000	+82.2% ∼ .177
Bayesian (Adj, μ)	4, 750	**.1148**	.184	.143	.082	+72.2% ∼ .007	+83.6% ∼ .044
Best	(2, 250	**.1232**	.228	.149	.077)	+82.7% ∼ .000	+76.0% ∼ .118
Coord. Level Ranking (Adj)	4	**.0651**	.184	.128	.071	+31.8% ∼ .319	+45.2% ∼ .058
TF-IDF (Adj)	4	**.0988**	.184	.159	.084	+63.4% ∼ .004	+72.7% ∼ .021
BM25 (Adj, b, k_1, k_3)	4, 0.05, 1.25, 1	**.1088**	.184	.163	.088	+60.2% ∼ .004	+66.2% ∼ .038
Best	(2, 0.30, 1.20, 200	**.1351**	.252	.163	.083)	+79.9% ∼ .000	+90.2% ∼ .047
Average of unbiased		**.0979**	.174	.134	.081	+57.3%	+59.2%
Average of best		**.1258**	.230	.161	.078	+81.5%	+77.5%

Table 32: Comparison of the *adj-only* ASR text representation with the *shot-only*, *video-only* and *story-only* representations for the different retrievals on the TRECVid 2003 collection.

TRECVID 2003	*Adj-Only*				*V. Shot-Only*	*V. Video-Only*	*V. Story-Only*
Retrieval Method	*Prm*	*MAP*	*P30*	*P100*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*
MLE (Adj)	2	**.0549**	.039	.001	+97.9% ∼ .087	+126.5% ∼ .150	-41.9% ∼ .002
Laplace (Adj)	2	**.0878**	.111	.056	+52.9% ∼ .058	+251.9% ∼ .004	-6.9% ∼ .681
Natural (Adj)	2	**.0863**	.117	.054	+45.3% ∼ .061	+266.0% ∼ .002	-27.7% ∼ .003
Lidstone (Adj, λ)	2, 0.34	**.0902**	.113	.056	+48.3% ∼ .196	+243.6% ∼ .004	-16.2% ∼ .198
Best	(10, 0.25	**.1087**	.145	.068)	+76.5% ∼ .116	+312.4% ∼ .000	-3.2% ∼ .272
Linear (Adj, α)	2, 0.01	**.0806**	.108	.054	+38.3% ∼ .152	+241.7% ∼ .000	-31.0% ∼ .003
Best	(4, 0.07	**.0892**	.095	.074)	+49.5% ∼ .163	+239.6% ∼ .000	-26.6% ∼ .014
Absolute Discounting (Adj, δ)	2, 0.89	**.0861**	.107	.059	+42.4% ∼ .214	+260.5% ∼ .000	-21.8% ∼ .058
Best	(4, 0.83	**.1004**	.103	.071)	+61.3% ∼ .238	+280.8% ∼ .000	-14.4% ∼ .046
Witten-Bell (Adj)	2	**.0927**	.136	.076	+26.9% ∼ .198	+298.5% ∼ .000	-28.1% ∼ .001
Jelinek-Mercer (Adj, λ)	2, 0.74	**.0906**	.127	.076	+24.0% ∼ .230	+258.3% ∼ .000	-30.6% ∼ .002
Best	(4, 0.02	**.0984**	.115	.086)	+33.7% ∼ .358	+285.3% ∼ .000	-26.2% ∼ .010
Absolute (Adj, δ)	2, 0.84	**.0929**	.129	.076	+28.1% ∼ .177	+262.1% ∼ .000	-26.1% ∼ .023
Best	(10, 0.85	**.1073**	.143	.076)	+46.1% ∼ .452	+306.4% ∼ .000	-17.9% ∼ .087
Dirichlet (Adj, μ)	2, 225	**.1008**	.128	.080	+30.2% ∼ .170	+297.7% ∼ .000	-17.4% ∼ .046
Best	(4, 750	**.1138**	.127	.088)	+46.5% ∼ .272	+344.3% ∼ .000	-9.1% ∼ .281
Bayesian (Adj, μ)	2, 250	**.1010**	.128	.080	+29.9% ∼ .156	+297.5% ∼ .000	-17.3% ∼ .044
Best	(4, 750	**.1138**	.127	.088)	+46.1% ∼ .272	+344.5% ∼ .000	-9.2% ∼ .246
Coord. Level Ranking (Adj)	2	**.0686**	.101	.057	+36.4% ∼ .206	+212.6% ∼ .001	-25.9% ∼ .611
TF-IDF (Adj)	2	**.0924**	.131	.067	+45.0% ∼ .079	+250.1% ∼ .000	-5.4% ∼ .823
BM25 (Adj, b, k_1, k_3)	2, 0.30, 1.20, 200	**.0969**	.137	.076	+29.9% ∼ .096	+282.4% ∼ .000	-16.9% ∼ .177
Best	(10, 0.00, 1.00, 300	**.1298**	.125	.076)	+65.6% ∼ .484	+385.6% ∼ .000	-2.3% ∼ .058
Average of unbiased		**.0873**	.115	.062	+41.1%	+253.5%	-22.4%
Average of best		**.1077**	.122	.078	+53.2%	+312.4%	-13.6%

Table 33: Comparison of the *adj-only* ASR text representation with the *shot-only* and *video-only* representations for the different retrievals on the TRECVid 2004 collection.

TRECVID 2004	*Adj-Only*				*V. Shot-Only*	*V. Video-Only*
Retrieval Method	*Prm*	*MAP*	*P30*	*P100*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*
MLE (Adj)	4	**.0294**	.028	.021	+30.8% ∼ .199	+766.8% ∼ .055
Laplace (Adj)	4	**.0404**	.090	.057	+13.9% ∼ .767	+871.1% ∼ .000
Natural (Adj)	4	**.0418**	.077	.057	+8.5% ∼ .464	+774.4% ∼ .001
Lidstone (Adj, λ)	4, 0.06	**.0433**	.091	.060	+21.3% ∼ .665	+878.7% ∼ .000
Best	(4, 0.02	**.0442**	.090	.060)	+23.7% ∼ .708	+794.4% ∼ .000
Linear (Adj, α)	4, 0.07	**.0419**	.083	.057	+3.9% ∼ .631	+751.9% ∼ .001
Best	(3, 0.56	**.0419**	.086	.056)	+3.7% ∼ .665	+748.1% ∼ .001
Absolute Discounting (Adj, δ)	4, 0.83	**.0425**	.083	.058	+27.7% ∼ .500	+895.9% ∼ .000
Best	(3, 0.54	**.0431**	.086	.059)	+6.7% ∼ .619	+773.6% ∼ .001
Witten-Bell (Adj)	4	**.0525**	.109	.070	+13.6% ∼ .572	+917.3% ∼ .000
Jelinek-Mercer (Adj, λ)	4, 0.02	**.0510**	.107	.067	+9.5% ∼ .608	+877.9% ∼ .000
Best	(3, 0.78	**.0545**	.113	.071)	+16.4% ∼ .381	+937.4% ∼ .000
Absolute (Adj, δ)	4, 0.85	**.0526**	.119	.070	+20.4% ∼ .596	+1,208.9% ∼ .000
Best	(3, 0.61	**.0543**	.112	.074)	+17.3% ∼ .346	+914.8% ∼ .001
Dirichlet (Adj, μ)	4, 750	**.0559**	.117	.070	+26.8% ∼ .560	+1,135.7% ∼ .000
Best	(2, 750	**.0583**	.116	.087)	+28.7% ∼ .243	+1,009.3% ∼ .000
Bayesian (Adj, μ)	4, 750	**.0557**	.119	.070	+27.5% ∼ .548	+1,129.6% ∼ .000
Best	(2, 750	**.0580**	.114	.087)	+29.0% ∼ .243	+1,007.8% ∼ .000
Coord. Level Ranking (Adj)	4	**.0291**	.062	.051	-17.8% ∼ .124	+1,741.3% ∼ .000
TF-IDF (Adj)	4	**.0563**	.117	.071	+27.2% ∼ .642	+946.5% ∼ .000
BM25 (Adj, b, k_1, k_3)	4, 0.05, 1.25, 1	**.0556**	.116	.071	+22.2% ∼ .488	+1,077.3% ∼ .000
Best	(2, 0.00, 1.25, 50	**.0586**	.113	.084)	+25.3% ∼ .331	+1,123.4% ∼ .000
Average of unbiased		**.0463**	.094	.061	+16.8%	+998.1%
Average of best		**.0516**	.104	.072	+18.9%	+913.6%

Table 34: Comparison of the *story-only* ASR text representation with the *shot-only*, *adj-only* and *video-only* representations for the different retrievals on the TRECVid 2003 collection.

TRECVID 2003	*Story-Only*				*V. Shot-Only*	*V. Adj-Only*	*V. Video-Only*
Retrieval Method	*Prm*	*MAP*	*P30*	*P100*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*
MLE		**.0945**	.091	.039	+240.9% ∼ .021	+72.2% ∼ .002	+290.0% ∼ .009
Laplace		**.0943**	.067	.072	+64.3% ∼ .431	+7.4% ∼ .681	+277.9% ∼ .000
Natural		**.1194**	.156	.101	+100.9% ∼ .044	+38.3% ∼ .003	+406.3% ∼ .000
Lidstone (λ)	0.02	**.1076**	.123	.096	+76.8% ∼ .132	+19.3% ∼ .198	+309.9% ∼ .000
Best	(0.01	**.1123**	.124	.100)	+82.3% ∼ .091	+3.3% ∼ .272	+326.1% ∼ .000
Linear (α)	0.97	**.1169**	.124	.101	+100.4% ∼ .019	+44.9% ∼ .003	+395.2% ∼ .000
Best	(0.67	**.1216**	.143	.100)	+103.8% ∼ .049	+36.3% ∼ .014	+362.9% ∼ .000
Absolute Discounting (δ)	0.94	**.1102**	.125	.092	+82.2% ∼ .058	+28.0% ∼ .058	+361.3% ∼ .000
Best	(0.40	**.1172**	.144	.095)	+88.3% ∼ .068	+16.8% ∼ .046	+344.7% ∼ .000
Witten-Bell		**.1290**	.145	.112	+76.4% ∼ .104	+39.0% ∼ .001	+454.1% ∼ .000
Jelinek-Mercer (λ)	0.76	**.1305**	.152	.110	+78.7% ∼ .090	+44.1% ∼ .002	+416.2% ∼ .000
Best	(0.92	**.1333**	.156	.113)	+81.2% ∼ .041	+35.5% ∼ .010	+422.0% ∼ .000
Absolute (δ)	0.80	**.1257**	.155	.111	+73.2% ∼ .085	+35.2% ∼ .023	+389.7% ∼ .000
Best	(0.64	**.1306**	.163	.109)	+77.8% ∼ .065	+21.7% ∼ .087	+394.7% ∼ .000
Dirichlet (μ)	75	**.1220**	.145	.109	+57.6% ∼ .180	+21.1% ∼ .046	+381.5% ∼ .000
Best	(50	**.1252**	.151	.108)	+61.3% ∼ .180	+10.1% ∼ .281	+389.0% ∼ .000
Bayesian (μ)	75	**.1222**	.145	.109	+57.1% ∼ .188	+21.0% ∼ .044	+381.0% ∼ .000
Best	(25	**.1253**	.152	.106)	+60.9% ∼ .163	+10.1% ∼ .246	+389.4% ∼ .000
Coord. Level Ranking		**.0925**	.071	.054	+84.0% ∼ .309	+34.9% ∼ .611	+321.6% ∼ .000
TF-IDF		**.0977**	.077	.070	+53.3% ∼ .473	+5.7% ∼ .823	+270.0% ∼ .000
BM25 (b, k_1, k_3)	0.40, 1.55, 100	**.1166**	.133	.106	+56.3% ∼ .159	+20.3% ∼ .177	+360.0% ∼ .000
Best	(1.00, 1.35, 20	**.1328**	.163	.106)	+69.5% ∼ .073	+2.3% ∼ .058	+396.9% ∼ .000
Average of unbiased		**.1128**	.122	.092	+85.9%	+30.8%	+358.2%
Average of best		**.1248**	.149	.105	+78.1%	+17.0%	+378.2%

Table 35: Statistical significance tests comparing standard retrieval models and language models using the *adj* structure on the TRECVid 2002 collection. Underlined entries are significant according to the one tailed Wilcoxon Sign Rank test with a 95% confidence ($p \leq 0.05$).

Retrieval Method	*MAP/%Dif*	*Wilcoxon Test Results*													
Absolute		-	>WB	>Dir	>Bay	>BM25	>AbsD	>JM	>Lid	>Lin	>TF	>Lap	>Nat	>CLR	>ML
(Abs)	0.1221		.043	.029	.029	.275	.018	.031	.017	.034	.129	.005	.000	.001	.000
Witten-Bell		<Abs	-	>Dir	>Bay	>BM25	>AbsD	>JM	>Lid	>Lin	>TF	>Lap	>Nat	>CLR	>ML
(WB)	0.1184/-3.1%	.043		.500	.500	.397	.358	.045	.179	.065	.391	.080	.008	.001	.000
Dirichlet		<Abs	<WB	-	>Bay	>BM25	>AbsD	>JM	>Lid	>Lin	>TF	>Lap	>Nat	>CLR	>ML
(Dir)	0.1166/-4.5%	.029	.500		.535	.500	.655	.251	.250	.249	.260	.031	.009	.001	.000
Bayesian		<Abs	<WB	<Dir	-	>BM25	>AbsD	>JM	>Lid	>Lin	>TF	>Lap	>Nat	>CLR	>ML
(Bay)	0.1148/-6.0%	.029	.500	.535		.707	.681	.314	.452	.411	.465	.073	.038	.001	.000
BM25		<Abs	<WB	<Dir	<Bay	-	>AbsD	>JM	>Lid	>Lin	>TF	>Lap	>Nat	>CLR	>ML
(BM25)	0.1088/-10.9%	.275	.397	.500	.707		.493	.359	.232	.148	.188	.018	.006	.000	.000
Absolute Discounting		<Abs	<WB	<Dir	<Bay	<BM25	-	>JM	>Lid	>Lin	>TF	>Lap	>Nat	>CLR	>ML
(AbsD)	0.1069/-12.5%	.018	.358	.655	.681	.493		.307	.397	.040	.468	.096	.008	.002	.000
Jelinek-Mercer		<Abs	<WB	<Dir	<Bay	<BM25	<AbsD	-	>Lid	>Lin	>TF	>Lap	>Nat	>CLR	>ML
(JM)	0.1069/-12.5%	.031	.045	.251	.314	.359	.307		.383	.092	.455	.206	.033	.001	.000
Lidstone		<Abs	<WB	<Dir	<Bay	<BM25	<AbsD	<JM	-	>Lin	>TF	>Lap	>Nat	>CLR	>ML
(Lid)	0.1022/-16.3%	.017	.179	.250	.452	.232	.397	.383		.263	.243	.031	.014	.001	.000
Linear		<Abs	<WB	<Dir	<Bay	<BM25	<AbsD	<JM	<Lid	-	>TF	>Lap	>Nat	>CLR	>ML
(Lin)	0.0995/-18.5%	.034	.065	.249	.411	.148	.040	.092	.263		.669	.515	.153	.006	.000
TF-IDF		<Abs	<WB	<Dir	<Bay	<BM25	<AbsD	<JM	<Lid	<Lin	-	>Lap	>Nat	>CLR	>ML
(TF)	0.0988/-19.1%	.129	.391	.260	.465	.188	.468	.455	.243	.669		.177	.094	.002	.000
Laplace		<Abs	<WB	<Dir	<Bay	<BM25	<AbsD	<JM	<Lid	<Lin	<TF	-	>Nat	>CLR	>ML
(Lap)	0.0978/-19.9%	.005	.080	.031	.073	.018	.096	.206	.031	.515	.177		.089	.006	.001
Natural		<Abs	<WB	<Dir	<Bay	<BM25	<AbsD	<JM	<Lid	<Lin	<TF	<Lap	-	>CLR	>ML
(Nat)	0.0859/-29.6%	.000	.008	.009	.038	.006	.008	.033	.014	.153	.094	.089		.003	.000
Coord. Level Ranking		<Abs	<WB	<Dir	<Bay	<BM25	<AbsD	<JM	<Lid	<Lin	<TF	<Lap	<Nat	-	>ML
(CLR)	0.0651/-46.7%	.001	.001	.001	.001	.000	.002	.001	.001	.006	.002	.006	.003		.002
MLE		<Abs	<WB	<Dir	<Bay	<BM25	<AbsD	<JM	<Lid	<Lin	<TF	<Lap	<Nat	<CLR	-
(ML)	0.0262/-78.6%	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.000	.002	

Table 36: Statistical significance tests comparing standard retrieval models and language models using the *adj* structure on the TRECVid 2003 collection. Underlined entries are significant according to the one tailed Wilcoxon Sign Rank test with a 95% confidence ($p \leq 0.05$).

Retrieval Method	*MAP/%Dif*	*Wilcoxon Test Results*													
Bayesian		-	>Dir	>BM25	>Abs	>WB	>TF	>JM	>Lid	>Lap	>Nat	>AbsD	>Lin	>CLR	>ML
(Bay)	0.1010		.321	.199	.055	.022	.264	.007	.009	.003	.020	.012	.000	.007	.000
Dirichlet		<Bay	-	>BM25	>Abs	>WB	>TF	>JM	>Lid	>Lap	>Nat	>AbsD	>Lin	>CLR	>ML
(Dir)	0.1008/-0.2%	.321		.356	.061	.025	.281	.006	.014	.003	.020	.012	.000	.007	.000
BM25		<Bay	<Dir	-	>Abs	>WB	>TF	>JM	>Lid	>Lap	>Nat	>AbsD	>Lin	>CLR	>ML
(BM25)	0.0969/-4.1%	.199	.356		.013	.144	.389	.066	.003	.001	.027	.004	.001	.016	.000
Absolute		<Bay	<Dir	<BM25	-	>WB	>TF	>JM	>Lid	>Lap	>Nat	>AbsD	>Lin	>CLR	>ML
(Abs)	0.0929/-8.0%	.055	.061	.013		.762	.700	.348	.079	.031	.064	.026	.003	.005	.000
Witten-Bell		<Bay	<Dir	<BM25	<Abs	-	>TF	>JM	>Lid	>Lap	>Nat	>AbsD	>Lin	>CLR	>ML
(WB)	0.0927/-8.2%	.022	.025	.144	.762		.726	.144	.191	.064	.091	.091	.001	.009	.000
TF-IDF		<Bay	<Dir	<BM25	<Abs	<WB	-	>JM	>Lid	>Lap	>Nat	>AbsD	>Lin	>CLR	>ML
(TF)	0.0924/-8.5%	.264	.281	.389	.700	.726		.198	.046	.011	.121	.046	.055	.034	.001
Jelinek-Mercer		<Bay	<Dir	<BM25	<Abs	<WB	<TF	-	>Lid	>Lap	>Nat	>AbsD	>Lin	>CLR	>ML
(JM)	0.0906/-10.3%	.007	.006	.066	.348	.144	.198		.324	.358	.163	.229	.014	.013	.000
Lidstone		<Bay	<Dir	<BM25	<Abs	<WB	<TF	<JM	-	>Lap	>Nat	>AbsD	>Lin	>CLR	>ML
(Lid)	0.0902/-10.7%	.009	.014	.003	.079	.191	.046	.324		.019	.274	.621	.079	.256	.000
Laplace		<Bay	<Dir	<BM25	<Abs	<WB	<TF	<JM	<Lid	-	>Nat	>AbsD	>Lin	>CLR	>ML
(Lap)	0.0878/-13.1%	.003	.003	.001	.031	.064	.011	.358	.019		.579	.823	.319	.366	.000
Natural		<Bay	<Dir	<BM25	<Abs	<WB	<TF	<JM	<Lid	<Lap	-	>AbsD	>Lin	>CLR	>ML
(Nat)	0.0863/-14.6%	.020	.020	.027	.064	.091	.121	.163	.274	.579		.855	.314	.238	.000
Absolute Discounting		<Bay	<Dir	<BM25	<Abs	<WB	<TF	<JM	<Lid	<Lap	<Nat	-	>Lin	>CLR	>ML
(AbsD)	0.0861/-14.8%	.012	.012	.004	.026	.091	.046	.229	.621	.823	.855		.030	.214	.000
Linear		<Bay	<Dir	<BM25	<Abs	<WB	<TF	<JM	<Lid	<Lap	<Nat	<AbsD	-	>CLR	>ML
(Lin)	0.0806/-20.2%	.000	.000	.001	.003	.001	.055	.014	.079	.319	.314	.030		.265	.000
Coord. Level Ranking		<Bay	<Dir	<BM25	<Abs	<WB	<TF	<JM	<Lid	<Lap	<Nat	<AbsD	<Lin	-	>ML
(CLR)	0.0686/-32.1%	.007	.007	.016	.005	.009	.034	.013	.256	.366	.238	.214	.265		.021
MLE		<Bay	<Dir	<BM25	<Abs	<WB	<TF	<JM	<Lid	<Lap	<Nat	<AbsD	<Lin	<CLR	-
(ML)	0.0549/-45.7%	.000	.000	.000	.000	.000	.001	.000	.000	.000	.000	.000	.000	.021	

Table 37: Statistical significance tests comparing standard retrieval models and language models using the *adj* structure on the TRECVid 2004 collection. Underlined entries are significant according to the one tailed Wilcoxon Sign Rank test with a 95% confidence ($p \leq 0.05$).

Retrieval Method	*MAP/%Dif*	*Wilcoxon Test Results*													
TF-IDF		-	>Dir	>Bay	>BM25	>Abs	>WB	>JM	>Lid	>AbsD	>Lin	>Nat	>Lap	>ML	>CLR
(TF)	0.0563		.494	.379	.146	.177	.213	.038	.002	.005	.002	.011	.000	.000	.000
Dirichlet		<TF	-	>Bay	>BM25	>Abs	>WB	>JM	>Lid	>AbsD	>Lin	>Nat	>Lap	>ML	>CLR
(Dir)	0.0559/-0.7%	.494		.605	.090	.046	.023	.008	.001	.003	.002	.006	.000	.000	.000
Bayesian		<TF	<Dir	-	>BM25	>Abs	>WB	>JM	>Lid	>AbsD	>Lin	>Nat	>Lap	>ML	>CLR
(Bay)	0.0557/-1.1%	.379	.605		.129	.070	.034	.016	.002	.004	.004	.012	.000	.000	.000
BM25		<TF	<Dir	<Bay	-	>Abs	>WB	>JM	>Lid	>AbsD	>Lin	>Nat	>Lap	>ML	>CLR
(BM25)	0.0556/-1.3%	.146	.090	.129		.153	.238	.037	.003	.002	.002	.008	.000	.000	.000
Absolute		<TF	<Dir	<Bay	<BM25	-	>WB	>JM	>Lid	>AbsD	>Lin	>Nat	>Lap	>ML	>CLR
(Abs)	0.0526/-6.5%	.177	.046	.070	.153		.705	.358	.002	.001	.002	.006	.000	.000	.001
Witten-Bell		<TF	<Dir	<Bay	<BM25	<Abs	-	>JM	>Lid	>AbsD	>Lin	>Nat	>Lap	>ML	>CLR
(WB)	0.0525/-6.7%	.213	.023	.034	.238	.705		.124	.003	.001	.000	.001	.000	.000	.000
Jelinek-Mercer		<TF	<Dir	<Bay	<BM25	<Abs	<WB	-	>Lid	>AbsD	>Lin	>Nat	>Lap	>ML	>CLR
(JM)	0.0510/-9.4%	.038	.008	.016	.037	.358	.124		.003	.002	.001	.003	.001	.000	.000
Lidstone		<TF	<Dir	<Bay	<BM25	<Abs	<WB	<JM	-	>AbsD	>Lin	>Nat	>Lap	>ML	>CLR
(Lid)	0.0433/-23.1%	.002	.001	.002	.003	.002	.003	.003		.079	.015	.106	.046	.000	.027
Absolute Discounting		<TF	<Dir	<Bay	<BM25	<Abs	<WB	<JM	<Lid	-	>Lin	>Nat	>Lap	>ML	>CLR
(AbsD)	0.0425/-24.6%	.005	.003	.004	.002	.001	.001	.002	.079		.272	.253	.161	.000	.047
Linear		<TF	<Dir	<Bay	<BM25	<Abs	<WB	<JM	<Lid	<AbsD	-	>Nat	>Lap	>ML	>CLR
(Lin)	0.0419/-25.6%	.002	.002	.004	.002	.002	.000	.001	.015	.272		.379	.777	.000	.331
Natural		<TF	<Dir	<Bay	<BM25	<Abs	<WB	<JM	<Lid	<AbsD	<Lin	-	>Lap	>ML	>CLR
(Nat)	0.0418/-25.8%	.011	.006	.012	.008	.006	.001	.003	.106	.253	.379		.676	.000	.572
Laplace		<TF	<Dir	<Bay	<BM25	<Abs	<WB	<JM	<Lid	<AbsD	<Lin	<Nat	-	>ML	>CLR
(Lap)	0.0404/-28.2%	.000	.000	.000	.000	.000	.000	.001	.046	.161	.777	.676		.000	.051
MLE		<TF	<Dir	<Bay	<BM25	<Abs	<WB	<JM	<Lid	<AbsD	<Lin	<Nat	<Lap	-	>CLR
(ML)	0.0294/-47.8%	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.998
Coord. Level Ranking		<TF	<Dir	<Bay	<BM25	<Abs	<WB	<JM	<Lid	<AbsD	<Lin	<Nat	<Lap	<ML	-
(CLR)	0.0291/-48.3%	.000	.000	.000	.000	.001	.000	.000	.027	.047	.331	.572	.051	.998	

Table 38: Statistical significance tests comparing standard retrieval models and language models using the *story* structure on the TRECVid 2004 collection. Underlined entries are significant according to the one tailed Wilcoxon Sign Rank test with a 95% confidence ($p \leq 0.05$).

Retrieval Method	*MAP/%Dif*	*Wilcoxon Test Results*													
Jelinek-Mercer		-	>WB	>Abs	>Bay	>Dir	>Nat	>Lin	>BM25	>AbsD	>Lid	>TF	>ML	>Lap	>CLR
(JM)	0.1305		.324	.065	.109	.102	.034	.034	.058	.004	.002	.001	.000	.000	.000
Witten-Bell		<JM	-	>Abs	>Bay	>Dir	>Nat	>Lin	>BM25	>AbsD	>Lid	>TF	>ML	>Lap	>CLR
(WB)	0.1290/-1.2%	.324		.115	.046	.037	.041	.075	.027	.007	.003	.000	.000	.000	.000
Absolute		<JM	<WB	-	>Bay	>Dir	>Nat	>Lin	>BM25	>AbsD	>Lid	>TF	>ML	>Lap	>CLR
(Abs)	0.1257/-3.7%	.065	.115		.629	.507	.404	.144	.063	.016	.013	.002	.004	.001	.001
Bayesian		<JM	<WB	<Abs	-	>Dir	>Nat	>Lin	>BM25	>AbsD	>Lid	>TF	>ML	>Lap	>CLR
(Bay)	0.1222/-6.4%	.109	.046	.629		.008	.274	.319	.131	.009	.003	.001	.002	.000	.000
Dirichlet		<JM	<WB	<Abs	<Bay	-	>Nat	>Lin	>BM25	>AbsD	>Lid	>TF	>ML	>Lap	>CLR
(Dir)	0.1220/-6.5%	.102	.037	.507	.008		.274	.331	.131	.009	.003	.001	.002	.000	.000
Natural		<JM	<WB	<Abs	<Bay	<Dir	-	>Lin	>BM25	>AbsD	>Lid	>TF	>ML	>Lap	>CLR
(Nat)	0.1194/-8.5%	.034	.041	.404	.274	.274		.366	.158	.013	.020	.008	.001	.001	.001
Linear		<JM	<WB	<Abs	<Bay	<Dir	<Nat	-	>BM25	>AbsD	>Lid	>TF	>ML	>Lap	>CLR
(Lin)	0.1169/-10.5%	.034	.075	.144	.319	.331	.366		.506	.256	.274	.052	.021	.021	.034
BM25		<JM	<WB	<Abs	<Bay	<Dir	<Nat	<Lin	-	>AbsD	>Lid	>TF	>ML	>Lap	>CLR
(BM25)	0.1166/-10.7%	.058	.027	.063	.131	.131	.158	.506		.033	.031	.003	.023	.003	.006
Absolute Discounting		<JM	<WB	<Abs	<Bay	<Dir	<Nat	<Lin	<BM25	-	>Lid	>TF	>ML	>Lap	>CLR
(AbsD)	0.1102/-15.6%	.004	.007	.016	.009	.009	.013	.256	.033		.354	.054	.021	.004	.041
Lidstone		<JM	<WB	<Abs	<Bay	<Dir	<Nat	<Lin	<BM25	<AbsD	-	>TF	>ML	>Lap	>CLR
(Lid)	0.1076/-17.6%	.002	.003	.013	.003	.003	.020	.274	.031	.354		.053	.010	.004	.006
TF-IDF		<JM	<WB	<Abs	<Bay	<Dir	<Nat	<Lin	<BM25	<AbsD	<Lid	-	>ML	>Lap	>CLR
(TF)	0.0977/-25.2%	.001	.000	.002	.001	.001	.008	.052	.003	.054	.053		.070	.153	.034
MLE		<JM	<WB	<Abs	<Bay	<Dir	<Nat	<Lin	<BM25	<AbsD	<Lid	<TF	-	>Lap	>CLR
(ML)	0.0945/-27.6%	.000	.000	.004	.002	.002	.001	.021	.023	.021	.010	.070		.911	.915
Laplace		<JM	<WB	<Abs	<Bay	<Dir	<Nat	<Lin	<BM25	<AbsD	<Lid	<TF	<ML	-	>CLR
(Lap)	0.0943/-27.8%	.000	.000	.001	.000	.000	.001	.021	.003	.004	.004	.153	.911		.548
Coord. Level Ranking		<JM	<WB	<Abs	<Bay	<Dir	<Nat	<Lin	<BM25	<AbsD	<Lid	<TF	<ML	<Lap	-
(CLR)	0.0925/-29.1%	.000	.000	.001	.000	.000	.001	.034	.006	.041	.006	.034	.915	.548	

Table 39: Statistical significance tests comparing hierarchical language models using the *shot+video* structure on the TRECVid 2002, TRECVid 2003 and TRECVid 2004 collections. Underlined entries are significant according to the one tailed Wilcoxon Sign Rank test with a 95% confidence ($p \leq 0.05$).

(a) TRECVid 2002

Retrieval Method	*MAP/%Dif*	*Wilcoxon Test Results*			
Jelinek-Mercer		-	>WB	>Abs	>Dir
(JM)	0.1335		.428	.224	.003
Witten-Bell		<JM	-	>Abs	>Dir
(WB)	0.1325/-0.7%	.428		.188	.025
Absolute		<JM	<WB	-	>Dir
(Abs)	0.1296/-2.9%	.224	.188		.010
Dirichlet		<JM	<WB	<Abs	-
(Dir)	0.1119/-16.2%	.003	.025	.010	

(b) TRECVid 2003

Retrieval Method	*MAP/%Dif*	*Wilcoxon Test Results*			
Jelinek-Mercer		-	>WB	>Abs	>Dir
(JM)	0.0945		.345	.013	.000
Witten-Bell		<JM	-	>Abs	>Dir
(WB)	0.0913/-3.4%	.345		.109	.001
Absolute		<JM	<WB	-	>Dir
(Abs)	0.0862/-8.8%	.013	.109		.002
Dirichlet		<JM	<WB	<Abs	-
(Dir)	0.0705/-25.5%	.000	.001	.002	

(c) TRECVid 2004

Retrieval Method	*MAP/%Dif*	*Wilcoxon Test Results*			
Jelinek-Mercer		-	>Dir	>Abs	>WB
(JM)	0.0506		.827	.047	.044
Dirichlet		<JM	-	>Abs	>WB
(Dir)	0.0489/-3.3%	.827		.181	.053
Absolute		<JM	<Dir	-	>WB
(Abs)	0.0488/-3.5%	.047	.181		.507
Witten-Bell		<JM	<Dir	<Abs	-
(WB)	0.0485/-4.0%	.044	.053	.507	

Table 40: Statistical significance tests comparing hierarchical language models using the *shot+adj+video* structure on the TRECVid 2002, TRECVid 2003 and TRECVid 2004 collections. Underlined entries are significant according to the one tailed Wilcoxon Sign Rank test with a 95% confidence ($p \leq 0.05$).

(a) TRECVid 2002

Retrieval Method	*MAP/%Dif*	*Wilcoxon Test Results*			
Jelinek-Mercer		-	>WB	>Abs	>Dir
(JM)	0.1605		.146	.173	.007
Witten-Bell		<JM	-	>Abs	>Dir
(WB)	0.1541/-4.0%	.146		.805	.173
Absolute		<JM	<WB	-	>Dir
(Abs)	0.1502/-6.5%	.173	.805		.342
Dirichlet		<JM	<WB	<Abs	-
(Dir)	0.1433/-10.7%	.007	.173	.342	

(b) TRECVid 2003

Retrieval Method	*MAP/%Dif*	*Wilcoxon Test Results*			
Jelinek-Mercer		-	>WB	>Abs	>Dir
(JM)	0.1405		.388	.001	.000
Witten-Bell		<JM	-	>Abs	>Dir
(WB)	0.1384/-1.5%	.388		.081	.003
Absolute		<JM	<WB	-	>Dir
(Abs)	0.1277/-9.1%	.001	.081		.013
Dirichlet		<JM	<WB	<Abs	-
(Dir)	0.1100/-21.7%	.000	.003	.013	

(c) TRECVid 2004

Retrieval Method	*MAP/%Dif*	*Wilcoxon Test Results*			
Jelinek-Mercer		-	>WB	>Dir	>Abs
(JM)	0.0686		.090	.608	.024
Witten-Bell		<JM	-	>Dir	>Abs
(WB)	0.0660/-3.7%	.090		.814	.252
Dirichlet		<JM	<WB	-	>Abs
(Dir)	0.0651/-5.0%	.608	.814		.158
Absolute		<JM	<WB	<Dir	-
(Abs)	0.0629/-8.2%	.024	.252	.158	

Table 41: Statistical significance tests comparing hierarchical language models using the *shot+story* structure on the TRECVid 2003 collections. Underlined entries are significant according to the one tailed Wilcoxon Sign Rank test with a 95% confidence ($p \leq 0.05$).

Retrieval Method	*MAP/%Dif*	*Wilcoxon Test Results*			
Jelinek-Mercer		-	>WB	>Abs	>Dir
(JM)	0.1551		.253	.094	.017
Witten-Bell		<JM	-	>Abs	>Dir
(WB)	0.1526/-1.6%	.253		.065	.029
Absolute		<JM	<WB	-	>Dir
(Abs)	0.1453/-6.3%	.094	.065		.272
Dirichlet		<JM	<WB	<Abs	-
(Dir)	0.1403/-9.6%	.017	.029	.272	

Table 42: Statistical significance tests comparing hierarchical language models using the *shot+adj+story* structure on the TRECVid 2003 collections. Underlined entries are significant according to the one tailed Wilcoxon Sign Rank test with a 95% confidence ($p \leq 0.05$).

Retrieval Method	*MAP/%Dif*	*Wilcoxon Test Results*			
Jelinek-Mercer		-	>WB	>Abs	>Dir
(JM)	0.1526		.335	.196	.041
Witten-Bell		<JM	-	>Abs	>Dir
(WB)	0.1499/-1.8%	.335		.229	.166
Absolute		<JM	<WB	-	>Dir
(Abs)	0.1467/-3.9%	.196	.229		.238
Dirichlet		<JM	<WB	<Abs	-
(Dir)	0.1424/-6.7%	.041	.166	.238	

APPENDIX C

ADDITIONAL TABLES FOR CHAPTER 6

Table 43: Comparison of retrieval models on the *global HSV 16x4x4* feature for the *TRECVid 2002* and *TRECVid 2003* search tasks.

HSV 16x4x4	*TRECVid 2002*					*TRECVid 2003*				
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Prm*	*MAP*	*P10*	*P30*	*P100*
MLE		**.0031**	.023	.016	.008		**.0007**	.007	.004	.002
Laplace		**.0156**	.034	.022	.015		**.0152**	.061	.049	.026
Natural		**.0166**	.033	.022	.017		**.0152**	.067	.047	.025
Lidstone	0.20	**.0158**	.034	.021	.016	0.01	**.0153**	.065	.048	.025
Best	(0.01	**.0162**	.034	.021	.017)	(0.20	**.0154**	.064	.048	.026)
Linear	0.01	**.0150**	.034	.023	.014	0.01	**.0151**	.060	.049	.026
Best	(0.01	**.0150**	.034	.023	.014)	(0.01	**.0151**	.060	.049	.026)
Absolute Discounting	0.10	**.0160**	.034	.021	.016	0.01	**.0152**	.064	.048	.025
Best	(0.01	**.0163**	.034	.022	.017)	(0.10	**.0153**	.065	.049	.025)
Witten-Bell		**.0158**	.034	.022	.015		**.0157**	.063	.050	.026
Jelinek Mercer	0.03	**.0151**	.033	.022	.014	0.20	**.0151**	.066	.047	.027
Best	(0.20	**.0155**	.034	.023	.014)	(0.03	**.0160**	.066	.049	.026)
Absolute	0.80	**.0158**	.034	.021	.016	0.01	**.0154**	.067	.048	.025
Best	(0.01	**.0163**	.034	.022	.017)	(0.80	**.0157**	.063	.049	.025)
Manhattan		**.0129**	.034	.022	.015		**.0122**	.058	.043	.024
Euclidean		**.0084**	.025	.017	.012		**.0094**	.049	.040	.022
Jensen-Shannon		**.0167**	.040	.023	.013		**.0138**	.061	.047	.025
Average of unbiased		**.0139**	.033	.021	.014		**.0132**	.057	.043	.023
Average of best		**.0159**	.034	.022	.016		**.0155**	.064	.049	.026

Table 44: Comparison of retrieval models on the regional *HSV 5x5 16x4x4* feature for the *TRECVid 2002* and *TRECVid 2003* search task.

HSV 5x5 16x4x4	*TRECVid 2002*					*V. HSV 16x4x4*	*TRECVid 2003*					*V. HSV 16x4x4*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ∼ Wilc.*	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ∼ Wilc.*
MLE		**.0000**	.000	.000	.000	-100.0% ∼ .000		**.0000**	.000	.000	.000	-100.0% ∼ .000
Laplace		**.0156**	.041	.028	.017	-0.3% ∼ .930		**.0280**	.094	.072	.042	+83.9% ∼ .000
Natural		**.0125**	.040	.025	.017	-24.4% ∼ .559		**.0247**	.083	.066	.039	+62.3% ∼ .046
Lidstone	1.55	**.0158**	.040	.027	.017	-0.1% ∼ .944	10.00	**.0270**	.089	.069	.042	+76.1% ∼ .000
Best	(10.00	**.0166**	.044	.028	.018)	+2.1% ∼ .410	(1.55	**.0280**	.094	.072	.042)	+82.4% ∼ .000
Linear	0.04	**.0156**	.041	.027	.017	+4.3% ∼ .024	0.20	**.0260**	.088	.068	.041	+72.0% ∼ .000
Best	(0.20	**.0164**	.044	.026	.017)	+9.5% ∼ .032	(0.04	**.0273**	.091	.070	.041)	+80.8% ∼ .001
Absolute Discounting	0.45	**.0147**	.038	.026	.018	-7.9% ∼ .707	0.95	**.0256**	.089	.068	.039	+68.4% ∼ .005
Best	(0.95	**.0151**	.042	.026	.017)	-7.2% ∼ .635	(0.45	**.0258**	.090	.068	.040)	+68.2% ∼ .020
Witten-Bell		**.0149**	.041	.026	.017	-5.8% ∼ .861		**.0259**	.088	.069	.040	+65.3% ∼ .020
Jelinek Mercer	0.02	**.0153**	.041	.026	.018	+1.7% ∼ .031	0.65	**.0238**	.080	.066	.037	+57.9% ∼ .000
Best	(0.65	**.0165**	.042	.028	.018)	+6.0% ∼ .207	(0.02	**.0276**	.094	.072	.042)	+72.9% ∼ .007
Absolute	0.30	**.0145**	.037	.026	.018	-8.1% ∼ .878	0.70	**.0258**	.089	.069	.040	+66.8% ∼ .010
Best	(0.70	**.0148**	.041	.026	.017)	-9.2% ∼ .791	(0.30	**.0259**	.089	.069	.040)	+65.2% ∼ .025
Manhattan		**.0174**	.042	.027	.020	+34.8% ∼ .078		**.0251**	.092	.071	.040	+105.9% ∼ .000
Euclidean		**.0057**	.027	.018	.012	-32.2% ∼ .174		**.0127**	.065	.051	.027	+34.6% ∼ .233
Jensen-Shannon		**.0187**	.047	.031	.020	+12.2% ∼ .083		**.0258**	.092	.071	.040	+87.0% ∼ .000
Average of unbiased		**.0134**	.036	.024	.016	-10.5%		**.0225**	.079	.062	.036	+56.7%
Average of best		**.0159**	.043	.027	.017	+0.3%		**.0269**	.092	.071	.041	+73.9%

Table 45: Comparison of the *indexing units* shots, sequence of adjacent shots, and videos with the HSV 80x1x1+1 colour representation for language models and visual retrieval models on the *TRECVid 2002* search task.

TRECVID 2002	*Video Only*					*V. Shot Only*	*Adj Only*					*V. Shot Only*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ∼ Wilc.*	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ∼ Wilc.*
MLE (Adj)		**.0169**	.005	.009	.021	+394.7% ∼ .211	1	**.0076**	.027	.024	.016	+124.3% ∼ .011
Best							(20	**.0239**	.025	.023	.021)	+600.3% ∼ .624
Laplace (Adj)		**.0169**	.005	.009	.021	+22.2% ∼ .627	1	**.0169**	.036	.026	.020	+22.5% ∼ .265
Best							(20	**.0239**	.025	.023	.021)	+73.0% ∼ .947
Natural (Adj)		**.0169**	.005	.009	.021	+18.6% ∼ .584	1	**.0176**	.038	.028	.021	+23.5% ∼ .309
Best							(20	**.0239**	.025	.023	.021)	+67.9% ∼ .930
Lidstone (λ, Adj)	0.00	**.0169**	.005	.009	.021	+21.7% ∼ .650	0.65, 1	**.0170**	.036	.027	.020	+22.3% ∼ .256
Best	(0.00	**.0169**	.005	.009	.021)	+19.3% ∼ .629	(0.00, 20	**.0239**	.025	.023	.021)	+68.9% ∼ .953
Linear (α, Adj)	0.00	**.0169**	.005	.009	.021	+44.2% ∼ .716	0.05, 1	**.0159**	.033	.025	.019	+35.5% ∼ .285
Best	(0.00	**.0169**	.005	.009	.021)	+44.2% ∼ .716	(0.00, 20	**.0239**	.025	.023	.021)	+104.2% ∼ .985
Absolute Discounting (δ, Adj)	0.00	**.0169**	.005	.009	.021	+20.2% ∼ .653	0.60, 1	**.0170**	.036	.026	.020	+20.8% ∼ .380
Best	(0.00	**.0169**	.005	.009	.021)	+20.2% ∼ .653	(1.00, 20	**.0239**	.025	.023	.021)	+70.5% ∼ .944
Witten-Bell (Adj)		**.0169**	.005	.009	.021	+20.6% ∼ .635	1	**.0170**	.037	.026	.020	+21.8% ∼ .259
Best							(20	**.0239**	.025	.023	.021)	+70.6% ∼ .946
Jelinek Mercer (λ, Adj)	0.60	**.0164**	.010	.010	.025	+32.1% ∼ .594	0.05, 1	**.0158**	.033	.026	.020	+27.2% ∼ .336
Best	(0.20	**.0167**	.005	.008	.022)	+34.1% ∼ .521	(0.25, 10	**.0227**	.019	.023	.018)	+82.7% ∼ .943
Absolute (δ, Adj)	0.00	**.0169**	.005	.009	.021	+19.9% ∼ .633	0.50, 1	**.0173**	.037	.026	.020	+22.7% ∼ .352
Best	(0.00	**.0169**	.005	.009	.021)	+18.9% ∼ .642	(0.00, 20	**.0239**	.025	.023	.021)	+68.3% ∼ .953
Manhattan (Adj)		**.0131**	.001	.005	.017	+23.3% ∼ .976	1	**.0134**	.026	.021	.018	+25.7% ∼ .952
Best							(20	**.0216**	.022	.021	.017)	+103.2% ∼ .998
Euclidean (Adj)		**.0128**	.001	.003	.017	+63.0% ∼ .983	1	**.0106**	.021	.018	.015	+34.7% ∼ .332
Best							(20	**.0171**	.015	.017	.015)	+117.7% ∼ .992
Jensen-Shannon (Adj)		**.0138**	.001	.006	.018	+1.7% ∼ .904	1	**.0171**	.032	.027	.020	+26.4% ∼ .655
Best							(20	**.0231**	.021	.021	.019)	+70.3% ∼ .994
Average of unbiased		**.0159**	.005	.008	.021	+56.8%		**.0153**	.033	.025	.019	+33.9%
Average of best		**.0168**	.005	.008	.021	+27.3%		**.0230**	.023	.022	.020	+124.8%

Table 46: Results for using the *indexing unit* sequence of adjacent shots with the HSV 80x1x1+1 colour representation for language models and visual retrieval models on the *TRECVid 2003* search task.

TRECVID 2003	*Adj Only*					*V. Shot Only*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ∼ Wilc.*
MLE (Adj)	20	**.0008**	.006	.006	.003	-84.9% ∼ .001
Best	(1	**.0017**	.009	.009	.009)	-67.9% ∼ .660
Laplace (Adj)	20	**.0008**	.006	.006	.003	-90.7% ∼ .000
Best	(1	**.0018**	.010	.009	.009)	-79.0% ∼ .023
Natural (Adj)	20	**.0008**	.006	.006	.003	-90.6% ∼ .000
Best	(1	**.0018**	.009	.009	.009)	-79.0% ∼ .048
Lidstone (λ, Adj)	0.00, 20	**.0008**	.006	.006	.003	-90.6% ∼ .000
Best	(0.65, 1	**.0018**	.010	.009	.009)	-79.0% ∼ .022
Linear (α, Adj)	0.00, 20	**.0008**	.006	.006	.003	-90.1% ∼ .000
Best	(0.05, 1	**.0018**	.009	.009	.008)	-77.7% ∼ .015
Absolute Discounting (δ, Adj)	1.00, 20	**.0008**	.006	.006	.003	-90.7% ∼ .000
Best	(0.60, 1	**.0018**	.010	.009	.008)	-79.0% ∼ .025
Witten-Bell (Adj)	20	**.0008**	.006	.006	.003	-90.6% ∼ .000
Best	(1	**.0018**	.010	.009	.008)	-79.0% ∼ .023
Jelinek Mercer (λ, Adj)	0.25, 10	**.0000**	.000	.000	.000	-100.0% ∼ .000
Best	(0.05, 1	**.0018**	.008	.008	.009)	-77.2% ∼ .005
Absolute (δ, Adj)	0.00, 20	**.0008**	.006	.006	.003	-90.6% ∼ .000
Best	(0.50, 1	**.0018**	.010	.009	.009)	-78.9% ∼ .029
Manhattan (Adj)	20	**.0006**	.007	.005	.003	-88.7% ∼ .000
Best	(1	**.0016**	.008	.007	.007)	-68.4% ∼ .000
Euclidean (Adj)	20	**.0009**	.004	.004	.003	-74.0% ∼ .000
Best	(1	**.0014**	.009	.007	.006)	-56.3% ∼ .001
Jensen-Shannon (Adj)	20	**.0008**	.007	.005	.003	-89.4% ∼ .000
Best	(1	**.0017**	.009	.008	.008)	-76.8% ∼ .001
Average of unbiased		**.0007**	.005	.005	.003	-89.3%
Average of best		**.0017**	.009	.009	.008	-74.9%

Table 47: Comparison of *global Canny 64+1 edge representations* with *Canny 4+1*, *Canny 16+1* and *Canny 32+1* representations for language models and standard visual retrieval models for the *TRECVid 2002* search task.

TRECVID 2002	*Canny 64*					*V. Canny 4*	*V. Canny 16*	*V. Canny 32*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr.*	*Impr.*	*Impr.*
MLE		**.0053**	.030	.023	.014	+50.9%	+24.5%	+2.9%
Laplace		**.0051**	.033	.022	.015	+49.1%	+23.7%	+2.8%
Natural		**.0053**	.034	.021	.014	+54.2%	+27.0%	+5.2%
Lidstone	7.00	**.0036**	.015	.017	.015	+5.2%	-13.5%	-13.2%
Best	(0.01	**.0055**	.034	.024	.014)	+58.1%	+24.8%	+7.5%
Linear	0.01	**.0052**	.033	.023	.015	+47.7%	+20.8%	-3.9%
Best	(0.00	**.0053**	.030	.023	.014)	+50.9%	+22.4%	-6.1%
Absolute Discounting	0.60	**.0053**	.029	.025	.015	+51.2%	+26.8%	+4.8%
Best	(2.00	**.0059**	.033	.021	.015)	+67.3%	+27.4%	+6.2%
Witten-Bell		**.0055**	.034	.024	.014	+58.6%	+31.5%	+7.6%
Jelinek-Mercer	0.05	**.0054**	.034	.025	.015	+56.2%	+23.5%	+4.4%
Best	(0.04	**.0056**	.034	.025	.015)	+52.2%	+13.5%	+5.2%
Absolute	0.60	**.0055**	.032	.026	.014	+58.0%	+31.2%	+7.1%
Best	(2.00	**.0057**	.030	.022	.014)	+64.0%	+21.6%	+2.1%
Manhattan		**.0034**	.016	.013	.012	+30.6%	+11.3%	-6.8%
Euclidean		**.0023**	.012	.011	.008	-5.7%	-16.8%	-6.3%
Jensen-Shannon		**.0053**	.034	.022	.015	+54.4%	+26.6%	+5.6%
Average of unbiased		**.0048**	.028	.021	.014	+42.5%	+18.0%	+0.9%
Average of best		**.0056**	.032	.023	.014	+58.5%	+21.9%	+3.0%

Table 48: Comparison of *global Canny 64+1 edge representations* with *Canny 4+1*, *Canny 16+1* and *Canny 32+1* representations for language models and standard visual retrieval models for the *TRECVid 2003* search task.

TRECVID 2003	*Canny 64*					*V. Canny 4*	*V. Canny 16*	*V. Canny 32*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr.*	*Impr.*	*Impr.*
MLE		**.0075**	.036	.030	.020	+46.3%	+18.9%	+10.3%
Laplace		**.0084**	.039	.031	.020	+60.5%	+28.5%	+16.8%
Natural		**.0082**	.039	.030	.020	+57.0%	+28.3%	+17.0%
Lidstone	0.01	**.0080**	.038	.030	.020	+54.4%	+21.1%	+14.4%
Best	(7.00	**.0086**	.049	.033	.022)	+62.6%	+29.5%	+12.3%
Linear	0.00	**.0075**	.036	.030	.020	+50.7%	+17.7%	+7.4%
Best	(0.01	**.0081**	.039	.033	.021)	+57.7%	+26.2%	+13.5%
Absolute Discounting	2.00	**.0070**	.036	.029	.019	+36.2%	+28.2%	+12.8%
Best	(0.60	**.0081**	.039	.031	.020)	+57.3%	+24.0%	+15.1%
Witten-Bell		**.0081**	.041	.031	.021	+54.6%	+25.5%	+15.6%
Jelinek-Mercer	0.04	**.0083**	.040	.032	.020	+73.5%	+100.9%	+71.1%
Best	(0.05	**.0083**	.040	.030	.019)	+61.0%	+27.2%	+13.6%
Absolute	2.00	**.0072**	.036	.030	.020	+42.8%	+26.0%	+11.8%
Best	(0.60	**.0080**	.039	.031	.020)	+54.6%	+24.0%	+14.7%
Manhattan		**.0053**	.030	.023	.017	+22.2%	+5.1%	+2.1%
Euclidean		**.0042**	.020	.018	.013	-0.9%	-3.1%	+1.9%
Jensen-Shannon		**.0083**	.039	.031	.020	+58.0%	+25.4%	+14.6%
Average of unbiased		**.0073**	.036	.029	.019	+46.3%	+26.9%	+16.3%
Average of best		**.0082**	.041	.032	.020	+58.7%	+26.2%	+13.8%

Table 49: Comparison of *regional Canny edge representations* (Canny 64+1 for 3x3 regions) using language models and standard visual retrieval models for the *TRECVid 2002* search task.

TRECVID 2002	*Canny 3x3 64+1*					*V. Canny 64*	*V. Canny 4x4 64+1*	*V. Canny 5x5 64+1*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr.*	*Impr.*	*Impr.*
MLE		**.0013**	.007	.008	.008	-74.7%	-33.2%	+387.6%
Laplace		**.0071**	.037	.020	.012	+37.2%	+14.9%	+17.8%
Natural		**.0075**	.040	.022	.014	+41.1%	+2.7%	+15.4%
Lidstone	4.00	**.0052**	.027	.015	.010	+44.0%	+14.3%	-7.8%
Best	(0.10	**.0079**	.041	.020	.014)	+43.9%	+7.1%	+19.5%
Linear	0.01	**.0069**	.033	.018	.013	+33.5%	+21.9%	+4.0%
Best	(0.01	**.0069**	.033	.018	.013)	+30.6%	+9.7%	-1.1%
Absolute Discounting	0.35	**.0075**	.037	.019	.014	+42.7%	+1.3%	+17.5%
Best	(0.85	**.0085**	.037	.018	.015)	+45.8%	+12.3%	+22.1%
Witten-Bell		**.0083**	.037	.022	.014	+50.1%	+11.0%	+23.7%
Jelinek-Mercer	0.40	**.0090**	.040	.021	.014	+65.8%	+15.7%	-2.2%
Best	(0.45	**.0091**	.038	.021	.015)	+61.0%	+5.1%	-4.7%
Absolute	0.30	**.0080**	.032	.022	.015	+44.5%	+9.1%	+23.1%
Best	(0.80	**.0098**	.037	.021	.016)	+71.0%	+34.7%	+51.3%
Manhattan		**.0057**	.026	.020	.012	+71.2%	-12.9%	+6.6%
Euclidean		**.0045**	.018	.013	.011	+94.0%	-5.4%	-13.5%
Jensen-Shannon		**.0072**	.034	.020	.012	+37.3%	+12.0%	+8.4%
Average of unbiased		**.0065**	.031	.018	.012	+40.6%	+4.3%	+40.0%
Average of best		**.0084**	.037	.020	.014	+50.5%	+13.8%	+17.4%

Table 50: Comparison of *regional Canny edge histogram representations* (Canny 64+1 for 5x5 regions) using language models and standard visual retrieval models for the *TRECVid 2003* search task.

TRECVID 2003	*Canny 5x5 64+1*					*V. Canny 64*	*V. Canny 3x3 64+1*	*V. Canny 4x4 64+1*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr.*	*Impr.*	*Impr.*
MLE		**.0006**	.011	.002	.000	-91.8%	-80.2%	-54.2%
Laplace		**.0118**	.049	.036	.022	+40.4%	+17.9%	+21.7%
Natural		**.0081**	.043	.034	.021	-0.9%	-3.4%	+2.4%
Lidstone	0.06	**.0093**	.044	.037	.023	+16.9%	+3.1%	+10.6%
Best	(1.55	**.0120**	.043	.033	.022)	+39.6%	+7.8%	+11.4%
Linear	0.06	**.0097**	.039	.028	.018	+28.5%	-4.8%	-2.9%
Best	(0.03	**.0121**	.047	.035	.023)	+49.3%	+19.3%	+17.3%
Absolute Discounting	0.95	**.0094**	.045	.035	.022	+33.7%	+3.1%	+6.1%
Best	(0.55	**.0102**	.047	.040	.023)	+25.3%	+7.9%	+12.1%
Witten-Bell		**.0088**	.042	.036	.022	+9.3%	+1.5%	+6.5%
Jelinek-Mercer	0.85	**.0105**	.036	.028	.024	+27.6%	-9.1%	+6.3%
Best	(0.55	**.0112**	.044	.036	.026)	+35.0%	-3.8%	+5.5%
Absolute	0.65	**.0083**	.040	.035	.022	+14.9%	+8.5%	+0.7%
Best	(0.40	**.0083**	.040	.035	.021)	+4.5%	+3.3%	+1.3%
Manhattan		**.0107**	.045	.033	.021	+103.3%	+21.4%	+15.3%
Euclidean		**.0079**	.038	.030	.017	+87.1%	+14.1%	+2.7%
Jensen-Shannon		**.0110**	.048	.039	.024	+32.8%	+9.7%	+12.6%
Average of unbiased		**.0089**	.040	.031	.020	+25.2%	-1.5%	+2.3%
Average of best		**.0108**	.044	.036	.023	+30.7%	+6.9%	+9.5%

Table 51: Comparison of *global DCT 4x4x4x4 representation* with the DCT 8x8x8 and DCT 3x3x3x3x3 representations using language models and standard visual retrieval models for the *TRECVid 2002* search task.

TRECVID 2002	*DCT 4x4x4x4*					*V. DCT 8x8x8*	*V. DCT 3x3x3x3x3*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr.*	*Impr.*
MLE		**.0002**	.001	.000	.000	null∞%	+3,700.8%
Laplace		**.0076**	.041	.031	.018	+23.6%	+90.2%
Natural		**.0069**	.029	.022	.015	+88.8%	+50.7%
Lidstone	1.30	**.0078**	.044	.032	.019	+25.0%	+92.1%
Best	(1.55	**.0079**	.042	.033	.019)	+4.2%	-15.7%
Linear	0.10	**.0074**	.037	.031	.019	+36.5%	+63.1%
Best	(0.15	**.0076**	.040	.030	.019)	+0.4%	+20.5%
Absolute Discounting	0.60	**.0070**	.034	.030	.018	+62.7%	+42.6%
Best	(0.20	**.0073**	.040	.024	.016)	+67.1%	-18.8%
Witten-Bell		**.0072**	.038	.026	.017	+46.3%	+44.0%
Jelinek-Mercer	0.30	**.0074**	.044	.028	.019	+23.5%	+50.6%
Best	(0.30	**.0074**	.044	.028	.019)	+16.7%	-18.8%
Absolute	0.60	**.0072**	.034	.022	.016	+74.5%	-20.5%
Best	(0.70	**.0072**	.036	.022	.017)	+59.2%	-26.6%
Manhattan		**.0078**	.034	.027	.020	+99.4%	+89.1%
Euclidean		**.0067**	.029	.024	.017	+210.8%	+59.3%
Jensen-Shannon		**.0070**	.044	.028	.017	+74.0%	+87.8%
Average of unbiased		**.0067**	.034	.025	.016	+69.6%	+362.5%
Average of best		**.0075**	.040	.027	.018	+29.5%	-11.9%

Table 52: Comparison of *global DCT 8x8x8 representation* with the DCT 3x3x3x3x3 representation using language models and standard visual retrieval models for the *TRECVid 2003* search task.

TRECVID 2003		*DCT 8x8x8*				*V. DCT 4x4x4x4*	*V. DCT 3x3x3x3x3*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr.*	*Impr.*
MLE		**.0000**	.000	.000	.000	-100.0%	-100.0%
Laplace		**.0115**	.065	.043	.025	+45.8%	+23.3%
Natural		**.0095**	.057	.035	.020	+57.8%	+18.1%
Lidstone	3.00	**.0100**	.052	.043	.026	+26.7%	+55.9%
Best	(1.20	**.0118**	.065	.044	.025)	+47.8%	+23.7%
Linear	0.30	**.0102**	.055	.038	.024	+44.3%	+19.9%
Best	(0.15	**.0111**	.063	.042	.024)	+55.4%	+16.3%
Absolute Discounting	0.70	**.0097**	.061	.039	.022	+61.2%	+44.5%
Best	(0.55	**.0099**	.061	.040	.022)	+61.3%	+7.6%
Witten-Bell		**.0114**	.066	.042	.023	+69.0%	+20.4%
Jelinek-Mercer	0.30	**.0124**	.068	.044	.026	+66.2%	+77.1%
Best	(0.35	**.0124**	.067	.043	.027)	+66.7%	+25.0%
Absolute	0.65	**.0105**	.063	.038	.020	+83.9%	+58.0%
Best	(0.95	**.0107**	.064	.040	.021)	+85.2%	+14.4%
Manhattan		**.0107**	.059	.041	.023	+45.1%	+33.1%
Euclidean		**.0068**	.045	.030	.018	+44.9%	+13.6%
Jensen-Shannon		**.0114**	.055	.040	.022	+55.8%	+22.5%
Average of unbiased		**.0095**	.054	.036	.021	+41.7%	+23.9%
Average of best		**.0112**	.064	.042	.024	+63.3%	+17.4%

Table 53: Comparison of language models and standard visual retrieval models using the *global DCT 3x3x3x3x3 representation* for the *TRECVid 2002* and *TRECVid 2003* search tasks.

DCT 3x3x3x3x3	*TRECVID 2002*					*TRECVID 2003*				
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Prm*	*MAP*	*P10*	*P30*	*P100*
MLE		**.0000**	.000	.000	.000		**.0001**	.001	.001	.000
Laplace		**.0040**	.016	.011	.012		**.0094**	.048	.031	.021
Natural		**.0046**	.015	.014	.011		**.0081**	.045	.032	.019
Lidstone	0.90	**.0041**	.016	.012	.012	0.02	**.0064**	.041	.027	.016
Best	(0.02	**.0094**	.018	.014	.010)	(0.90	**.0095**	.046	.031	.021)
Linear	0.03	**.0046**	.018	.015	.011	0.01	**.0085**	.045	.035	.019
Best	(0.01	**.0063**	.016	.015	.011)	(0.03	**.0096**	.046	.033	.020)
Absolute Discounting	0.25	**.0049**	.018	.014	.012	0.01	**.0067**	.041	.028	.017
Best	(0.01	**.0090**	.016	.014	.009)	(0.25	**.0092**	.046	.033	.019)
Witten-Bell		**.0050**	.015	.016	.013		**.0095**	.052	.035	.021
Jelinek-Mercer	0.15	**.0049**	.014	.014	.013	0.01	**.0070**	.041	.029	.017
Best	(0.01	**.0092**	.019	.014	.010)	(0.15	**.0099**	.055	.034	.021)
Absolute	0.85	**.0090**	.015	.016	.013	0.05	**.0066**	.042	.028	.017
Best	(0.05	**.0098**	.018	.013	.010)	(0.85	**.0094**	.054	.035	.021)
Manhattan		**.0041**	.021	.014	.013		**.0081**	.041	.027	.019
Euclidean		**.0042**	.019	.016	.012		**.0060**	.034	.024	.015
Jensen-Shannon		**.0037**	.018	.013	.012		**.0093**	.047	.033	.019
Average of unbiased		**.0044**	.015	.013	.011		**.0071**	.040	.027	.017
Average of best		**.0088**	.018	.014	.010		**.0095**	.050	.033	.020

Table 54: Comparison of the *5x5 regional DCT 3x3x3x3x3 representations* with the non-regional texture representation using language models and standard visual retrieval models for the *TRECVid 2002* search task.

TRECVID 2002		*DCT 5x5 3x3x3x3x3*				*V. DCT 3x3x3x3x3*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr.*
MLE		**.0000**	.000	.000	.000	-100.0%
Laplace		**.0062**	.026	.017	.015	+54.3%
Natural		**.0061**	.030	.020	.016	+32.0%
Lidstone	0.07	**.0062**	.026	.018	.015	+52.4%
Best	(3.00	**.0063**	.025	.019	.013)	-32.4%
Linear	0.25	**.0062**	.025	.018	.015	+37.2%
Best	(0.45	**.0064**	.026	.016	.014)	+1.2%
Absolute Discounting	0.15	**.0062**	.027	.020	.016	+26.0%
Best	(0.65	**.0063**	.025	.020	.015)	-30.5%
Witten-Bell		**.0068**	.030	.020	.015	+35.8%
Jelinek-Mercer	0.85	**.0073**	.023	.018	.015	+47.5%
Best	(0.65	**.0074**	.025	.019	.015)	-19.5%
Absolute	0.50	**.0069**	.030	.021	.015	-23.2%
Best	(0.90	**.0079**	.027	.020	.016)	-20.0%
Manhattan		**.0063**	.026	.018	.015	+53.3%
Euclidean		**.0040**	.018	.017	.014	-5.9%
Jensen-Shannon		**.0063**	.025	.016	.015	+69.5%
Average of unbiased		**.0057**	.024	.017	.014	+23.2%
Average of best		**.0069**	.025	.019	.015	-20.2%

Table 55: Comparison of the *5x5 regional DCT 3x3x3x3x3 texture representations* with the non-regional and regional 3x3 and 4x4 representations using language models and standard visual retrieval models for the *TRECVid 2003* search task.

TRECVID 2003		*DCT 5x5 3x3x3x3x3*				*V. DCT 3x3x3x3x3*	*V. DCT 3x3 3x3x3x3x3*	*V. DCT 4x4 3x3x3x3x3*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr.*	*Impr.*	*Impr.*
MLE		**.0000**	.000	.000	.000	-100.0%	N/A	N/A
Laplace		**.0156**	.058	.046	.032	+66.6%	+13.1%	+1.7%
Natural		**.0168**	.067	.055	.034	+109.3%	+14.6%	+2.5%
Lidstone	3.00	**.0118**	.045	.034	.026	+83.8%	-25.6%	-33.4%
Best	(0.07	**.0191**	.067	.056	.037)	+100.7%	+20.1%	+5.5%
Linear	0.45	**.0181**	.067	.052	.036	+112.3%	+18.7%	+3.2%
Best	(0.25	**.0190**	.067	.054	.036)	+98.5%	+18.7%	+5.8%
Absolute Discounting	0.65	**.0169**	.062	.053	.032	+151.9%	+18.2%	+5.4%
Best	(0.15	**.0171**	.065	.057	.034)	+85.5%	+19.2%	+5.5%
Witten-Bell		**.0207**	.077	.063	.040	+119.4%	+19.4%	+5.0%
Jelinek-Mercer	0.65	**.0226**	.082	.064	.040	+224.3%	+35.9%	+9.9%
Best	(0.85	**.0231**	.080	.066	.041)	+132.5%	+28.6%	+9.1%
Absolute	0.90	**.0175**	.073	.057	.035	+164.2%	+12.3%	+9.7%
Best	(0.50	**.0189**	.078	.060	.037)	+101.6%	+17.5%	+5.1%
Manhattan		**.0172**	.064	.054	.035	+113.7%	+16.5%	+10.4%
Euclidean		**.0107**	.052	.039	.024	+77.9%	+21.1%	+1.0%
Jensen-Shannon		**.0183**	.064	.053	.034	+96.8%	+18.3%	+8.3%
Average of unbiased		**.0155**	.059	.048	.031	+101.7%	+14.8%	+2.2%
Average of best		**.0194**	.071	.059	.037	+103.8%	+20.8%	+6.2%

Table 56: Comparison of unbiased retrieval models for the TRECVid 2002 search tasks 5x5 regional HSV colour, Canny edge and DCT texture. Underlined entries are significant according to the Wilcoxon Sign Rank test.

Retrieval Method	*MAP*	*Wilcoxon Test Results*												
Jelinek-Mercer		-	>JD	>Man	>Int	>Lin	>WB	>Abs	>Lap	>Lid	>AbsD	>Nat	>Euc	>ML
(JM)	0.0108		.222	.021	.021	.006	.183	.079	.030	.015	.019	.003	.000	.000
Jensen-Shannon		<JM	-	>Man	>Int	>Lin	>WB	>Abs	>Lap	>Lid	>AbsD	>Nat	>Euc	>ML
(JD)	0.0106	.222		.005	.006	.005	.540	.261	.002	.001	.264	.053	.000	.000
Manhattan		<JM	<JD	-	>Int	>Lin	>WB	>Abs	>Lap	>Lid	>AbsD	>Nat	>Euc	>ML
(Man)	0.0097	.021	.005		.839	.282	.933	.738	.661	.217	.697	.174	.000	.000
Linear		<JM	<JD	<Man	<Int	-	>WB	>Abs	>Lap	>Lid	>AbsD	>Nat	>Euc	>ML
(Lin)	0.0095	.006	.005	.282	.353		.914	.688	.985	.941	.584	.116	.000	.000
Witten-Bell		<JM	<JD	<Man	<Int	<Lin	-	>Abs	>Lap	>Lid	>AbsD	>Nat	>Euc	>ML
(WB)	0.0095	.183	.540	.933	.933	.914		.031	.204	.064	.009	.000	.000	.000
Absolute		<JM	<JD	<Man	<Int	<Lin	<WB	-	>Lap	>Lid	>AbsD	>Nat	>Euc	>ML
(Abs)	0.0093	.079	.261	.738	.739	.688	.031		.604	.385	.174	.007	.000	.000
Laplace		<JM	<JD	<Man	<Int	<Lin	<WB	<Abs	-	>Lid	>AbsD	>Nat	>Euc	>ML
(Lap)	0.0092	.030	.002	.661	.660	.985	.204	.604		.075	.346	.069	.000	.000
Lidstone		<JM	<JD	<Man	<Int	<Lin	<WB	<Abs	<Lap	-	>AbsD	>Nat	>Euc	>ML
(Lid)	0.0092	.015	.001	.217	.254	.941	.064	.385	.075		.607	.078	.000	.000
Absolute Discounting		<JM	<JD	<Man	<Int	<Lin	<WB	<Abs	<Lap	<Lid	-	>Nat	>Euc	>ML
(AbsD)	0.0091	.019	.264	.697	.709	.584	.009	.174	.346	.607		.028	.000	.000
Natural		<JM	<JD	<Man	<Int	<Lin	<WB	<Abs	<Lap	<Lid	<AbsD	-	>Euc	>ML
(Nat)	0.0084	.003	.053	.174	.177	.116	.000	.007	.069	.078	.028		.000	.000
Euclidean		<JM	<JD	<Man	<Int	<Lin	<WB	<Abs	<Lap	<Lid	<AbsD	<Nat	-	>ML
(Euc)	0.0050	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000
MLE		<JM	<JD	<Man	<Int	<Lin	<WB	<Abs	<Lap	<Lid	<AbsD	<Nat	<Euc	-
(ML)	0.0001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	

Table 57: Comparison of unbiased retrieval models for the TRECVid 2003 search tasks 5x5 regional HSV colour, Canny edge and DCT texture. Underlined entries are significant according to the Wilcoxon Sign Rank test.

Retrieval Method	*MAP*	*Wilcoxon Test Results*												
Jelinek-Mercer		-	>WB	>Lap	>JD	>Lin	>Int	>Man	>AbsD	>Abs	>Nat	>Lid	>Euc	>ML
(JM)	0.0190		.001	.001	.579	.001	.338	.337	.001	.000	.000	.000	.000	.000
Witten-Bell		<JM	-	>Lap	>JD	>Lin	>Int	>Man	>AbsD	>Abs	>Nat	>Lid	>Euc	>ML
(WB)	0.0185	.001		.820	.986	.718	.941	.941	.070	.000	.000	.637	.000	.000
Laplace		<JM	<WB	-	>JD	>Lin	>Int	>Man	>AbsD	>Abs	>Nat	>Lid	>Euc	>ML
(Lap)	0.0185	.001	.820		1.000	.694	.997	.996	.005	.002	.000	.002	.000	.000
Jensen-Shannon		<JM	<WB	<Lap	-	>Lin	>Int	>Man	>AbsD	>Abs	>Nat	>Lid	>Euc	>ML
(JD)	0.0183	.579	.986	1.000		.000	.013	.012	.000	.000	.000	.000	.000	.000
Linear		<JM	<WB	<Lap	<JD	-	>Int	>Man	>AbsD	>Abs	>Nat	>Lid	>Euc	>ML
(Lin)	0.0179	.001	.718	.694	.000		.997	.997	.008	.004	.000	.044	.000	.000
Manhattan		<JM	<WB	<Lap	<JD	<Lin	<Int	-	>AbsD	>Abs	>Nat	>Lid	>Euc	>ML
(Man)	0.0177	.337	.941	.996	.012	.997	.476		.000	.001	.000	.000	.000	.000
Absolute Discounting		<JM	<WB	<Lap	<JD	<Lin	<Int	<Man	-	>Abs	>Nat	>Lid	>Euc	>ML
(AbsD)	0.0173	.001	.070	.005	.000	.008	.000	.000		.029	.000	.527	.000	.000
Absolute		<JM	<WB	<Lap	<JD	<Lin	<Int	<Man	<AbsD	-	>Nat	>Lid	>Euc	>ML
(Abs)	0.0172	.000	.000	.002	.000	.004	.001	.001	.029		.074	.971	.000	.000
Natural		<JM	<WB	<Lap	<JD	<Lin	<Int	<Man	<AbsD	<Abs	-	>Lid	>Euc	>ML
(Nat)	0.0166	.000	.000	.000	.000	.000	.000	.000	.000	.074		1.000	.000	.000
Lidstone		<JM	<WB	<Lap	<JD	<Lin	<Int	<Man	<AbsD	<Abs	<Nat	-	>Euc	>ML
(Lid)	0.0163	.000	.637	.002	.000	.044	.000	.000	.527	.971	1.000		.000	.000
Euclidean		<JM	<WB	<Lap	<JD	<Lin	<Int	<Man	<AbsD	<Abs	<Nat	<Lid	-	>ML
(Euc)	0.0104	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000
MLE		<JM	<WB	<Lap	<JD	<Lin	<Int	<Man	<AbsD	<Abs	<Nat	<Lid	<Euc	-
(ML)	0.0002	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	

Table 58: Comparison of unbiased retrieval models for the TRECVid 2004 search tasks 5x5 regional HSV colour, Canny edge and DCT texture. Underlined entries are significant according to the Wilcoxon Sign Rank test.

Retrieval Method	*MAP*	*Wilcoxon Test Results*											
Absolute Discounting		-	>JD	>JM	>Man	>WB	>Abs	**>Nat**	>Lin	>Lid	>Lap	>Euc	>ML
(AbsD)	0.0077		.390	.131	.249	.361	.353	.024	.214	.110	.046	.000	.000
Jensen-Shannon		<AbsD	-	>JM	>Man	>WB	>Abs	>Nat	>Lin	>Lid	>Lap	>Euc	>ML
(JD)	0.0076	.390		.018	.033	.576	.600	.428	.180	.115	.001	.000	.000
Jelinek-Mercer		<AbsD	<JD	-	>Man	>WB	>Abs	>Nat	>Lin	>Lid	>Lap	>Euc	>ML
(JM)	0.0076	.131	.018		.911	.913	.837	.519	.981	.981	.366	.002	.000
Manhattan		<AbsD	<JD	<JM	-	>WB	>Abs	>Nat	>Lin	>Lid	>Lap	>Euc	>ML
(Man)	0.0075	.249	.033	.911		.556	.765	.329	.376	.147	.006	.000	.000
Witten-Bell		<AbsD	<JD	<JM	<Man	-	>Abs	**>Nat**	>Lin	>Lid	>Lap	>Euc	>ML
(WB)	0.0074	.361	.576	.913	.556		.469	.043	.399	.206	.047	.000	.000
Absolute		<AbsD	<JD	<JM	<Man	<WB	-	>Nat	>Lin	>Lid	>Lap	>Euc	>ML
(Abs)	0.0073	.353	.600	.837	.765	.469		.117	.420	.272	.061	.000	.000
Natural		**<AbsD**	<JD	<JM	<Man	**<WB**	<Abs	-	>Lin	>Lid	>Lap	**>Euc**	>ML
(Nat)	0.0072	.024	.428	.519	.329	.043	.117		.630	.262	.176	.003	.000
Linear		<AbsD	<JD	<JM	<Man	<WB	<Abs	<Nat	-	>Lid	>Lap	**>Euc**	>ML
(Lin)	0.0071	.214	.180	.981	.376	.399	.420	.630		.169	.052	.001	.000
Lidstone		<AbsD	<JD	<JM	<Man	<WB	<Abs	<Nat	<Lin	-	>Lap	**>Euc**	>ML
(Lid)	0.0070	.110	.115	.981	.147	.206	.272	.262	.169		.062	.002	.000
Laplace		**<AbsD**	**<JD**	<JM	**<Man**	**<WB**	<Abs	<Nat	<Lin	<Lid	-	>Euc	>ML
(Lap)	0.0061	.046	.001	.366	.006	.047	.061	.176	.052	.062		.092	.000
Euclidean		**<AbsD**	**<JD**	**<JM**	**<Man**	**<WB**	**<Abs**	**<Nat**	**<Lin**	**<Lid**	<Lap	-	>ML
(Euc)	0.0054	.000	.000	.002	.000	.000	.000	.003	.001	.002	.092		.000
MLE		**<AbsD**	**<JD**	**<JM**	**<Man**	**<WB**	**<Abs**	**<Nat**	**<Lin**	**<Lid**	**<Lap**	**<Euc**	-
(ML)	0.0000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	

APPENDIX D

ADDITIONAL TABLES FOR CHAPTER 7

Table 59: Comparison of the *Vis-CombWtScore* fusion of the colour, edge and texture results with the colour-only results and the other fusion methods *Vis-CombJointPr* and *Vis-CombWtRank* on the TRECVid 2002 collection.

TRECVid 2002	*Vis-CombWtScore*					*V. Colour-Only*	*V. Vis-CombJointPr*	*V. Vis-CombWtRank*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ~ Wilc.*	*Impr. ~ Wilc.*	*Impr. ~ Wilc.*
Laplace	0.600, 0.150, 0.250	**.0160**	.048	.032	.020	+2.8% ~ .050	-3.2% ~ .984	+26.1% ~ .000
Best	(0.700, 0.250, 0.050	**.0176**	.044	.029	.019)	+12.8% ~ .004	+6.2% ~ .019	-0.1% ~ .919
Lidstone	0.600, 0.150, 0.250	**.0164**	.053	.032	.020	+3.8% ~ .008	-0.3% ~ .995	+3.8% ~ .008
Best	(0.700, 0.250, 0.050	**.0172**	.048	.028	.018)	+3.6% ~ .089	+2.1% ~ .029	-2.2% ~ .956
Linear	0.600, 0.200, 0.200	**.0158**	.052	.033	.020	+1.2% ~ .036	-2.9% ~ .984	-5.0% ~ .975
Best	(0.700, 0.250, 0.050	**.0174**	.047	.028	.019)	+6.2% ~ .030	+3.0% ~ .100	-2.4% ~ .945
Absolute Discounting	0.600, 0.150, 0.250	**.0137**	.055	.033	.018	-6.6% ~ .993	-9.5% ~ .975	+10.2% ~ .006
Best	(0.800, 0.200, 0.000	**.0164**	.041	.030	.018)	+8.0% ~ .001	+6.3% ~ .001	+5.2% ~ .001
Witten-Bell	0.500, 0.200, 0.300	**.0139**	.056	.033	.019	-6.8% ~ .986	-10.3% ~ .963	+11.2% ~ .002
Best	(0.900, 0.100, 0.000	**.0157**	.042	.027	.018)	+5.1% ~ .000	+1.2% ~ .002	+1.6% ~ .001
Jelinek-Mercer	0.500, 0.150, 0.350	**.0143**	.052	.032	.020	-7.0% ~ .994	-8.6% ~ .984	+29.2% ~ .164
Best	(0.800, 0.150, 0.050	**.0170**	.047	.028	.018)	+3.5% ~ .015	+1.9% ~ .056	+2.8% ~ .009
Absolute	0.600, 0.150, 0.250	**.0134**	.055	.032	.018	-7.8% ~ .993	-10.0% ~ .988	+10.9% ~ .009
Best	(0.900, 0.100, 0.000	**.0153**	.042	.027	.017)	+3.1% ~ .002	+0.4% ~ .004	+3.1% ~ .003
Manhattan	0.600, 0.150, 0.250	**.0175**	.056	.035	.020	+0.5% ~ .039	N/A	-0.5% ~ .949
Best	(0.600, 0.200, 0.200	**.0185**	.058	.037	.020)	+6.0% ~ .032	N/A	+4.9% ~ .034
Jensen Shannon	0.600, 0.150, 0.250	**.0181**	.056	.036	.021	-3.1% ~ .999	N/A	+47.7% ~ .742
Best	(0.600, 0.250, 0.150	**.0197**	.058	.036	.021)	+5.5% ~ .012	N/A	-0.7% ~ .964
Average of unbiased		**.0155**	.054	.033	.020	-2.6%	-6.4%	+14.8%
Average of best		**.0172**	.047	.030	.019	+6.0%	+3.0%	+1.4%

Table 60: Comparison of the *Vis-CombWtScore* fusion of the colour, edge and texture results with the colour-only results and the other fusion methods *Vis-CombJointPr* and *Vis-CombWtRank* on the TRECVid 2003 collection.

TRECVid 2003	*Vis-CombWtScore*					*V. Colour-Only*	*V. Vis-CombJointPr*	*V. Vis-CombWtRank*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ~ Wilc.*	*Impr. ~ Wilc.*	*Impr. ~ Wilc.*
Laplace	0.700, 0.250, 0.050	**.0282**	.093	.075	.044	+0.9% ~ .037	-0.4% ~ .891	+2.4% ~ .001
Best	(0.600, 0.150, 0.250	**.0292**	.090	.074	.046)	+4.5% ~ .001	+3.2% ~ .000	+3.7% ~ .011
Lidstone	0.700, 0.250, 0.050	**.0275**	.090	.072	.045	+2.1% ~ .030	-2.6% ~ .234	+3.3% ~ .007
Best	(0.600, 0.150, 0.250	**.0296**	.092	.077	.047)	+5.6% ~ .001	+4.0% ~ .000	+5.6% ~ .001
Linear	0.700, 0.250, 0.050	**.0268**	.088	.070	.043	+2.9% ~ .017	+0.9% ~ .101	+3.2% ~ .001
Best	(0.600, 0.200, 0.200	**.0292**	.092	.076	.046)	+7.0% ~ .000	+5.7% ~ .000	+7.0% ~ .000
Absolute Discounting	0.800, 0.200, 0.000	**.0260**	.087	.071	.041	+1.8% ~ .005	+0.3% ~ .005	+2.8% ~ .004
Best	(0.600, 0.150, 0.250	**.0276**	.088	.072	.043)	+7.1% ~ .000	+6.0% ~ .000	+5.5% ~ .000
Witten-Bell	0.900, 0.100, 0.000	**.0262**	.088	.070	.041	+1.1% ~ .053	-0.3% ~ .479	+5.3% ~ .016
Best	(0.500, 0.200, 0.300	**.0290**	.093	.077	.047)	+12.1% ~ .000	+10.6% ~ .000	+9.6% ~ .000
Jelinek-Mercer	0.800, 0.150, 0.050	**.0252**	.083	.068	.039	+5.8% ~ .000	+3.2% ~ .000	+9.4% ~ .000
Best	(0.500, 0.150, 0.350	**.0324**	.102	.082	.051)	+17.5% ~ .000	+16.6% ~ .000	+4.5% ~ .271
Absolute	0.900, 0.100, 0.000	**.0261**	.089	.069	.041	+1.2% ~ .063	-0.2% ~ .537	+1.2% ~ .063
Best	(0.600, 0.150, 0.250	**.0286**	.090	.071	.044)	+10.4% ~ .000	+9.0% ~ .000	+7.9% ~ .000
Manhattan	0.600, 0.200, 0.200	**.0273**	.095	.078	.046	+8.4% ~ .000	N/A	+8.1% ~ .000
Best	(0.600, 0.150, 0.250	**.0275**	.097	.078	.046)	+9.2% ~ .000	N/A	+8.9% ~ .001
Jensen Shannon	0.600, 0.250, 0.150	**.0271**	.092	.078	.046	+4.9% ~ .002	N/A	+6.6% ~ .000
Best	(0.600, 0.150, 0.250	**.0279**	.092	.077	.047)	+7.9% ~ .000	N/A	+7.1% ~ .891
Average of unbiased		**.0267**	.089	.072	.043	+3.2%	+0.1%	+4.7%
Average of best		**.0290**	.093	.076	.047	+9.0%	+7.9%	+6.6%

Table 61: Comparison of the *Vis-CombWtScore* fusion of the colour, edge and texture results with the colour-only results and the other fusion methods *Vis-CombJointPr* and *Vis-CombWtRank* on the TRECVid 2004 collection.

TRECVid 2004	*Vis-CombWtScore*					*V. Colour-Only*	*V. Vis-CombJointPr*	*V. Vis-CombWtRank*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ~ Wilc.*	*Impr. ~ Wilc.*	*Impr. ~ Wilc.*
Laplace	0.600, 0.150, 0.250	**.0096**	.036	.028	.015	+7.4% ~ .915	+0.9% ~ .760	+9.4% ~ .834
Best	(0.500, 0.400, 0.100	**.0109**	.037	.029	.016)	+22.1% ~ .979	+14.8% ~ .923	+0.6% ~ .771
Lidstone	0.600, 0.150, 0.250	**.0111**	.041	.029	.016	+25.9% ~ .192	+17.9% ~ .186	+25.9% ~ .192
Best	(0.400, 0.250, 0.350	**.0125**	.043	.030	.016)	+41.7% ~ .311	+32.7% ~ .111	-2.6% ~ .247
Linear	0.600, 0.200, 0.200	**.0111**	.040	.030	.015	+24.7% ~ .162	+17.4% ~ .266	+19.6% ~ .145
Best	(0.400, 0.250, 0.350	**.0125**	.045	.030	.016)	+40.2% ~ .213	+32.1% ~ .078	-3.1% ~ .385
Absolute Discounting	0.600, 0.150, 0.250	**.0114**	.044	.028	.015	+30.4% ~ .080	+26.5% ~ .076	+14.4% ~ .019
Best	(0.300, 0.300, 0.400	**.0132**	.041	.031	.017)	+51.4% ~ .145	+47.0% ~ .080	-0.0% ~ .112
Witten-Bell	0.500, 0.200, 0.300	**.0119**	.041	.027	.017	+37.4% ~ .026	+35.4% ~ .094	+23.6% ~ .107
Best	(0.400, 0.200, 0.400	**.0126**	.038	.028	.017)	+45.9% ~ .085	+43.8% ~ .050	-1.1% ~ .156
Jelinek-Mercer	0.500, 0.150, 0.350	**.0113**	.040	.030	.016	+29.5% ~ .514	+21.9% ~ .484	+7.8% ~ .623
Best	(0.400, 0.400, 0.200	**.0135**	.041	.029	.018)	+53.7% ~ .644	+44.6% ~ .458	-0.4% ~ .010
Absolute	0.600, 0.150, 0.250	**.0110**	.043	.027	.015	+26.4% ~ .032	+27.2% ~ .136	+14.7% ~ .037
Best	(0.400, 0.150, 0.450	**.0126**	.036	.028	.016)	+45.2% ~ .132	+46.0% ~ .085	-1.5% ~ .218
Manhattan	0.600, 0.150, 0.250	**.0114**	.041	.026	.016	+43.6% ~ .063	N/A	+15.4% ~ .127
Best	(0.400, 0.400, 0.200	**.0133**	.048	.029	.016)	+67.8% ~ .489	N/A	-9.9% ~ .023
Jensen Shannon	0.600, 0.150, 0.250	**.0120**	.043	.030	.017	+31.3% ~ .031	N/A	-12.7% ~ .090
Best	(0.400, 0.300, 0.300	**.0132**	.046	.032	.018)	+44.5% ~ .076	N/A	-5.8% ~ .316
Average of unbiased		**.0112**	.041	.028	.016	+28.5%	+21.0%	+13.1%
Average of best		**.0127**	.042	.029	.017	+45.8%	+37.3%	-2.7%

Table 62: Comparison of unbiased retrieval models for the TRECVid 2002 search tasks using the *Vis-CombWtScore* fusion method for combining colour, edge and texture results.

Ret. Method	*MAP/%Dif*	*Wilcoxon Test Results*								
Jensen-Shannon		-	>Man	>Lid	>Lap	>Lin	>JM	>WB	>AbsD	>Abs
(JD)	.0181		.002	.077	.047	.050	.267	.091	.002	.030
Manhattan		<JD	-	>Lid	>Lap	>Lin	>JM	>WB	>AbsD	>Abs
(Man)	.0175/-3.3%	.002		.323	.306	.271	.476	.249	.065	.133
Lidstone		<JD	<Man	-	>Lap	>Lin	>JM	>WB	>AbsD	>Abs
(Lid)	.0164/-9.4%	.077	.323		.382	.056	.814	.281	.010	.124
Laplace		<JD	<Man	<Lid	-	>Lin	>JM	>WB	>AbsD	>Abs
(Lap)	.0160/-11.6%	.047	.306	.382		.100	.896	.385	.015	.026
Linear		<JD	<Man	<Lid	<Lap	-	>JM	>WB	>AbsD	>Abs
(Lin)	.0158/-12.7%	.050	.271	.056	.100		.858	.413	.025	.115
Jelinek-Mercer		<JD	<Man	<Lid	<Lap	<Lin	-	>WB	>AbsD	>Abs
(JM)	.0143/-21.0%	.267	.476	.814	.896	.858		.052	.018	.033
Witten-Bell		<JD	<Man	<Lid	<Lap	<Lin	<JM	-	>AbsD	>Abs
(WB)	.0139/-23.2%	.091	.249	.281	.385	.413	.052		.004	.009
Absolute Discounting		<JD	<Man	<Lid	<Lap	<Lin	<JM	<WB	-	>Abs
(AbsD)	.0137/-24.3%	.002	.065	.010	.015	.025	.018	.004		.772
Absolute		<JD	<Man	<Lid	<Lap	<Lin	<JM	<WB	<AbsD	-
(Abs)	.0134/-26.0%	.030	.133	.124	.026	.115	.033	.009	.772	

Table 63: Comparison of unbiased retrieval models for the TRECVid 2003 search tasks using the *Vis-CombWtScore* fusion method for combining colour, edge and texture results.

Ret. Method	*MAP/%Dif*	*Wilcoxon Test Results*								
Laplace		-	>Lid	>Man	>JD	>Lin	>WB	>Abs	>AbsD	>JM
(Lap)	.0282		.796	.996	.999	.621	.000	.000	.000	.736
Lidstone		<Lap	-	>Man	>JD	>Lin	>WB	>Abs	>AbsD	>JM
(Lid)	.0275/-1.5%	.796		.997	.999	.128	.000	.000	.000	.592
Manhattan		<Lap	<Lid	-	>JD	>Lin	>WB	>Abs	>AbsD	>JM
(Man)	.0273/-3.2%	.996	.997		.601	.000	.000	.000	.000	.000
Jensen-Shannon		<Lap	<Lid	<Man	-	>Lin	>WB	>Abs	>AbsD	>JM
(JD)	.0271/-3.9%	.999	.999	.601		.000	.000	.000	.000	.000
Linear		<Lap	<Lid	<Man	<JD	-	>WB	>Abs	>AbsD	>JM
(Lin)	.0268/-5.0%	.621	.128	.000	.000		.000	.000	.000	.401
Witten-Bell		<Lap	<Lid	<Man	<JD	<Lin	-	>Abs	>AbsD	>JM
(WB)	.0262/-7.0%	.000	.000	.000	.000	.000		.023	.995	.997
Absolute		<Lap	<Lid	<Man	<JD	<Lin	<WB	-	>AbsD	>JM
(Abs)	.0261/-7.4%	.000	.000	.000	.000	.000	.023		.998	.997
Absolute Discounting		<Lap	<Lid	<Man	<JD	<Lin	<WB	<Abs	-	>JM
(AbsD)	.0260/-7.8%	.000	.000	.000	.000	.000	.995	.998		.986
Jelinek-Mercer		<Lap	<Lid	<Man	<JD	<Lin	<WB	<Abs	<AbsD	-
(JM)	.0252/-10.6%	.736	.592	.000	.000	.401	.997	.997	.986	

Table 64: Comparison of unbiased retrieval models for the TRECVid 2004 search tasks using the *Vis-CombWtScore* fusion method for combining colour, edge and texture results.

Ret. Method	*MAP/%Dif*	*Wilcoxon Test Results*								
Jensen-Shannon		-	>WB	>Man	>AbsD	>JM	>Lid	>Lin	>Abs	>Lap
(JD)	.0120		.360	.038	.124	.198	.147	.148	.123	.034
Witten-Bell		<JD	-	>Man	>AbsD	>JM	>Lid	>Lin	>Abs	>Lap
(WB)	.0119/-0.8%	.360		.381	.045	.068	.242	.251	.075	.093
Manhattan		<JD	<WB	-	>AbsD	>JM	>Lid	>Lin	>Abs	>Lap
(Man)	.0114/-5.0%	.038	.381		.275	.358	.580	.500	.279	.280
Absolute Discounting		<JD	<WB	<Man	-	>JM	>Lid	>Lin	>Abs	>Lap
(AbsD)	.0114/-5.0%	.124	.045	.275		.439	.569	.333	.456	.343
Jelinek-Mercer		<JD	<WB	<Man	<AbsD	-	>Lid	>Lin	>Abs	>Lap
(JM)	.0113/-5.8%	.198	.068	.358	.439		.909	.825	.441	.440
Lidstone		<JD	<WB	<Man	<AbsD	<JM	-	>Lin	>Abs	>Lap
(Lid)	.0111/-7.5%	.147	.242	.580	.569	.909		.230	.429	.010
Linear		<JD	<WB	<Man	<AbsD	<JM	<Lid	-	>Abs	>Lap
(Lin)	.0111/-7.5%	.148	.251	.500	.333	.825	.230		.303	.033
Absolute		<JD	<WB	<Man	<AbsD	<JM	<Lid	<Lin	-	>Lap
(Abs)	.0110/-8.3%	.123	.075	.279	.456	.441	.429	.303		.395
Laplace		<JD	<WB	<Man	<AbsD	<JM	<Lid	<Lin	<Abs	-
(Lap)	.0096/-20.0%	.034	.093	.280	.343	.440	.010	.033	.395	

Table 65: Comparison of the *VisExs-CombMaxScore* fusion method with the other fusion methods *VisExs-CombJointPr*, *VisExs-CombRank*, *VisExs-CombScore* and *VisExs-CombMaxRank* for combining the visual examples' results of the different retrieval models on the TRECVid 2002 collection.

TRECVid 2002	*VisExs-CombMaxScore*				*V. VisExs-CombJointPr*	*V. VisExs-CombRank*	*V. VisExs-CombScore*	*V. VisExs-CombMaxRank*
Retrieval Method	*MAP*	*P10*	*P30*	*P100*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*
Laplace	**.0203**	.080	.052	.025	+211.9% ∼ .009	+34.2% ∼ .186	+15.5% ∼ .035	-0.0% ∼ .012
Best	**(.0229**	.068	.039	.024)	+251.9% ∼ .027	+31.0% ∼ .405	+15.0% ∼ .092	-0.1% ∼ .035
Lidstone	**.0206**	.080	.053	.027	+211.4% ∼ .006	+33.4% ∼ .210	+29.1% ∼ .022	-2.0% ∼ .023
Best	**(.0248**	.072	.044	.024)	+250.9% ∼ .012	+58.6% ∼ .263	+37.0% ∼ .054	-0.3% ∼ .001
Linear	**.0207**	.080	.052	.026	+224.8% ∼ .009	+39.3% ∼ .247	+22.5% ∼ .068	-4.6% ∼ .008
Best	**(.0214**	.068	.035	.023)	+214.9% ∼ .011	+38.4% ∼ .260	+21.2% ∼ .079	+0.1% ∼ .956
Absolute Discounting	**.0190**	.080	.052	.026	+234.4% ∼ .008	+52.6% ∼ .261	+39.7% ∼ .015	-0.1% ∼ .070
Best	**(.0236**	.072	.040	.022)	+305.3% ∼ .037	+71.4% ∼ .491	+30.9% ∼ .310	+0.5% ∼ .968
Witten-Bell	**.0209**	.084	.059	.028	+235.9% ∼ .001	+63.2% ∼ .068	+15.8% ∼ .116	+0.5% ∼ .603
Best	**(.0221**	.068	.040	.024)	+255.1% ∼ .029	+65.7% ∼ .261	+35.7% ∼ .074	-1.3% ∼ .048
Jelinek-Mercer	**.0212**	.080	.052	.032	+199.5% ∼ .001	+48.4% ∼ .137	+18.3% ∼ .153	+0.2% ∼ .566
Best	**(.0244**	.064	.043	.025)	+91.2% ∼ .005	+51.4% ∼ .148	+12.9% ∼ .006	+1.6% ∼ .980
Absolute	**.0202**	.080	.055	.026	+239.4% ∼ .001	+57.5% ∼ .092	+31.7% ∼ .016	-0.2% ∼ .044
Best	**(.0217**	.064	.043	.023)	+258.2% ∼ .042	+68.2% ∼ .261	+36.9% ∼ .096	+2.6% ∼ .833
Manhattan	**.0275**	.084	.055	.033	N/A	+87.8% ∼ .109	+42.1% ∼ .157	+0.1% ∼ .939
Best	**(.0276**	.076	.051	.033)	N/A	+72.4% ∼ .058	+40.3% ∼ .148	+0.1% ∼ .951
Jensen Shannon	**.0286**	.092	.059	.031	N/A	+81.2% ∼ .004	+47.6% ∼ .026	+3.1% ∼ .749
Best	**(.0306**	.084	.055	.032)	N/A	+76.7% ∼ .008	+47.0% ∼ .046	+4.1% ∼ .915
Average of unbiased	**.0221**	.082	.054	.028	+222.5%	+55.3%	+29.1%	-0.3%
Average of best	**.0243**	.071	.043	.025	+232.5%	+59.3%	+30.8%	+0.8%

Table 66: Comparison of the *VisExs-CombMaxScore* fusion method with the other fusion methods *VisExs-CombJointPr*, *VisExs-CombRank*, *VisExs-CombScore* and *VisExs-CombMaxRank* for combining the visual examples' results of the different retrieval models on the TRECVid 2003 collection.

TRECVid 2003	*VisExs-CombMaxScore*				*V. VisExs-CombJointPr*	*V. VisExs-CombRank*	*V. VisExs-CombScore*	*V. VisExs-CombMaxRank*
Retrieval Method	*MAP*	*P10*	*P30*	*P100*	*Impr. ~ Wilc.*	*Impr. ~ Wilc.*	*Impr. ~ Wilc.*	*Impr. ~ Wilc.*
Laplace	**.0463**	.108	.093	.060	+19.9% ~ .018	+17.6% ~ .213	+12.7% ~ .106	+0.3% ~ .660
Best	**(.0483**	.116	.092	.062)	+25.2% ~ .007	+16.1% ~ .233	+13.2% ~ .061	+1.1% ~ .223
Lidstone	**.0443**	.120	.099	.058	+4.3% ~ .262	+8.2% ~ .130	+7.5% ~ .094	+0.4% ~ .961
Best	**(.0499**	.116	.091	.062)	+25.7% ~ .007	+18.7% ~ .112	+17.4% ~ .173	+0.7% ~ .206
Linear	**.0468**	.104	.092	.057	+17.1% ~ .043	+15.3% ~ .099	+14.8% ~ .012	+0.6% ~ .921
Best	**(.0498**	.116	.093	.064)	+35.8% ~ .000	+23.3% ~ .079	+19.8% ~ .084	+0.5% ~ .354
Absolute Discounting	**.0367**	.100	.089	.051	+18.8% ~ .000	+7.2% ~ .068	+6.8% ~ .167	+1.3% ~ .528
Best	**(.0446**	.112	.084	.056)	+48.6% ~ .000	+24.0% ~ .034	+22.0% ~ .031	+1.2% ~ .515
Witten-Bell	**.0373**	.096	.088	.051	+21.4% ~ .002	+9.1% ~ .045	+6.6% ~ .086	+1.4% ~ .261
Best	**(.0499**	.124	.096	.066)	+62.5% ~ .000	+21.7% ~ .033	+21.4% ~ .084	+1.7% ~ .656
Jelinek-Mercer	**.0420**	.096	.081	.059	-25.5% ~ .173	+12.7% ~ .072	+2.5% ~ .112	+2.1% ~ .656
Best	**(.0554**	.120	.097	.068)	+50.3% ~ .000	+14.8% ~ .011	+15.4% ~ .177	+0.2% ~ .301
Absolute	**.0358**	.092	.088	.050	+18.0% ~ .002	+5.6% ~ .043	+4.1% ~ .063	+1.2% ~ .372
Best	**(.0469**	.120	.091	.058)	+55.4% ~ .000	+29.8% ~ .015	+27.0% ~ .054	+0.7% ~ .578
Manhattan	**.0444**	.128	.097	.061	N/A	+4.1% ~ .227	+8.3% ~ .384	+0.6% ~ .863
Best	**(.0460**	.128	.096	.062)	N/A	+6.3% ~ .296	+10.4% ~ .416	+0.4% ~ .908
Jensen Shannon	**.0444**	.128	.099	.064	N/A	+7.4% ~ .367	+13.4% ~ .233	+0.1% ~ .884
Best	**(.0467**	.124	.091	.062)	N/A	+9.7% ~ .715	+16.2% ~ .177	+1.4% ~ .830
Average of unbiased	**.0420**	.108	.092	.057	+10.6%	+9.7%	+8.5%	+0.9%
Average of best	**.0486**	.120	.092	.062	+43.4%	+18.3%	+18.1%	+0.9%

Table 67: Comparison of the *VisExs-CombMaxScore* fusion method with the other fusion methods *VisExs-CombJointPr*, *VisExs-CombRank*, *VisExs-CombScore* and *VisExs-CombMaxRank* for combining the visual examples' results of the different retrieval models on the TRECVid 2004 collection.

TRECVid 2004	*VisExs-CombMaxScore*				*V. VisExs-CombJointPr*	*V. VisExs-CombRank*	*V. VisExs-CombScore*	*V. VisExs-CombMaxRank*
Retrieval Method	*MAP*	*P10*	*P30*	*P100*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*
Laplace	**.0210**	.065	.042	.027	+407.6% ∼ .186	+23.8% ∼ .301	+12.3% ∼ .288	+0.6% ∼ .962
Best	**(.0296**	.109	.068	.035)	+616.9% ∼ .106	+24.6% ∼ .084	+54.6% ∼ .116	+2.3% ∼ .934
Lidstone	**.0250**	.091	.057	.033	+513.5% ∼ .169	+8.5% ∼ .145	+7.7% ∼ .179	-0.1% ∼ .010
Best	**(.0265**	.104	.058	.033)	+549.9% ∼ .161	+13.1% ∼ .047	+9.2% ∼ .084	-1.2% ∼ .013
Linear	**.0264**	.096	.062	.035	+563.2% ∼ .125	+16.8% ∼ .096	+21.3% ∼ .061	+0.0% ∼ .867
Best	**(.0268**	.096	.059	.033)	+573.5% ∼ .054	+17.4% ∼ .029	+20.5% ∼ .041	-0.8% ∼ .022
Absolute Discounting	**.0319**	.117	.077	.037	+920.9% ∼ .066	+89.8% ∼ .047	+52.9% ∼ .020	+1.0% ∼ .628
Best	**(.0344**	.122	.081	.039)	+1,003.4% ∼ .020	+103.5% ∼ .002	+63.2% ∼ .003	+1.6% ∼ .988
Witten-Bell	**.0311**	.122	.077	.039	+912.6% ∼ .021	+67.8% ∼ .023	+30.8% ∼ .125	+0.3% ∼ .977
Best	**(.0319**	.130	.074	.036)	+935.7% ∼ .010	+73.0% ∼ .051	+25.7% ∼ .494	+0.1% ∼ .977
Jelinek-Mercer	**.0222**	.083	.045	.027	+524.8% ∼ .131	+6.6% ∼ .084	+4.1% ∼ .125	-0.8% ∼ .000
Best	**(.0267**	.083	.052	.029)	+653.2% ∼ .616	+25.8% ∼ .888	+11.2% ∼ .900	-0.3% ∼ .017
Absolute	**.0315**	.117	.077	.038	+963.5% ∼ .044	+95.6% ∼ .008	+47.2% ∼ .072	+1.0% ∼ .910
Best	**(.0324**	.135	.075	.039)	+996.8% ∼ .011	+101.7% ∼ .001	+49.9% ∼ .001	+1.5% ∼ .982
Manhattan	**.0284**	.104	.064	.035	N/A	+34.2% ∼ .023	+13.1% ∼ .455	+0.6% ∼ .986
Best	**(.0362**	.130	.083	.038)	N/A	+26.0% ∼ .102	+28.8% ∼ .576	+0.8% ∼ .934
Jensen Shannon	**.0286**	.109	.067	.039	N/A	+21.9% ∼ .021	+5.3% ∼ .391	+0.3% ∼ .946
Best	**(.0320**	.113	.081	.038)	N/A	+29.3% ∼ .027	+8.0% ∼ .506	+2.0% ∼ .971
Average of unbiased	**.0273**	.100	.063	.034	+686.6%	+40.5%	+21.6%	+0.3%
Average of best	**.0307**	.114	.070	.035	+761.3%	+46.1%	+30.1%	+0.7%

Table 68: Comparison of unbiased retrieval models for the TRECVid 2002 search tasks using the *VisExs-CombMaxScore* fusion method for combining visual examples.

Ret. Method	*MAP/%Dif*	*Wilcoxon Test Results*								
Jensen-Shannon		-	>Man	>JM	>WB	>Lin	>Lid	>Lap	>Abs	>AbsD
(JD)	0.0286		.049	.121	.104	.002	.004	.005	.173	.006
Manhattan		<JD	-	>JM	>WB	>Lin	>Lid	>Lap	>Abs	>AbsD
(Man)	0.0275/-3.8%	.049		.256	.099	.006	.013	.018	.189	.021
Jelinek-Mercer		<JD	<Man	-	>WB	>Lin	>Lid	>Lap	>Abs	>AbsD
(JM)	0.0212/-25.9%	.121	.256		.516	.014	.012	.006	.334	.024
Witten-Bell		<JD	<Man	<JM	-	>Lin	>Lid	>Lap	>Abs	>AbsD
(WB)	0.0209/-26.9%	.104	.099	.516		.073	.081	.060	.047	.022
Linear		<JD	<Man	<JM	<WB	-	>Lid	>Lap	>Abs	>AbsD
(Lin)	0.0207/-27.6%	.002	.006	.014	.073		.669	.589	.849	.479
Lidstone		<JD	<Man	<JM	<WB	<Lin	-	>Lap	>Abs	>AbsD
(Lid)	0.0206/-28.0%	.004	.013	.012	.081	.669		.295	.633	.145
Laplace		<JD	<Man	<JM	<WB	<Lin	<Lid	-	>Abs	>AbsD
(Lap)	0.0203/-29.0%	.005	.018	.006	.060	.589	.295		.646	.275
Absolute		<JD	<Man	<JM	<WB	<Lin	<Lid	<Lap	-	>AbsD
(Abs)	0.0202/-29.4%	.173	.189	.334	.047	.849	.633	.646		.023
Absolute Discounting		<JD	<Man	<JM	<WB	<Lin	<Lid	<Lap	<Abs	-
(AbsD)	0.0190/-33.6%	.006	.021	.024	.022	.479	.145	.275	.023	

Table 69: Comparison of unbiased retrieval models for the TRECVid 2003 search tasks using the *VisExs-CombMaxScore* fusion method for combining visual examples.

Ret. Method	*MAP/%Dif*	*Wilcoxon Test Results*								
Linear		-	>Lap	>JD	>Man	>Lid	>JM	>WB	>AbsD	>Abs
(Lin)	0.0468		.052	.848	.744	.557	.274	.003	.000	.002
Laplace		<Lin	-	>JD	>Man	>Lid	>JM	>WB	>AbsD	>Abs
(Lap)	0.0463/-1.1%	.052		.964	.914	.708	.697	.005	.002	.003
Jensen-Shannon		<Lin	<Lap	-	>Man	>Lid	>JM	>WB	>AbsD	>Abs
(JD)	0.0444/-5.1%	.848	.964		.468	.090	.274	.010	.001	.007
Manhattan		<Lin	<Lap	<JD	-	>Lid	>JM	>WB	>AbsD	>Abs
(Man)	0.0444/-5.1%	.744	.914	.468		.101	.252	.010	.002	.006
Lidstone		<Lin	<Lap	<JD	<Man	-	>JM	>WB	>AbsD	>Abs
(Lid)	0.0443/-5.3%	.557	.708	.090	.101		.558	.002	.003	.002
Jelinek-Mercer		<Lin	<Lap	<JD	<Man	<Lid	-	>WB	>AbsD	>Abs
(JM)	0.0420/-10.3%	.274	.697	.274	.252	.558		.022	.012	.022
Witten-Bell		<Lin	<Lap	<JD	<Man	<Lid	<JM	-	>AbsD	>Abs
(WB)	0.0373/-20.3%	.003	.005	.010	.010	.002	.022		.516	.175
Absolute Discounting		<Lin	<Lap	<JD	<Man	<Lid	<JM	<WB	-	>Abs
(AbsD)	0.0367/-21.6%	.000	.002	.001	.002	.003	.012	.516		.210
Absolute		<Lin	<Lap	<JD	<Man	<Lid	<JM	<WB	<AbsD	-
(Abs)	0.0358/-23.5%	.002	.003	.007	.006	.002	.022	.175	.210	

Table 70: Comparison of unbiased retrieval models for the TRECVid 2004 search tasks using the *VisExs-CombMaxScore* fusion method for combining visual examples.

Ret. Method	*MAP/%Dif*	*Wilcoxon Test Results*								
Absolute Discounting		-	>Abs	>WB	>JD	>Man	>Lin	>Lid	>JM	>Lap
(AbsD)	0.0319		.558	.704	.916	.847	.404	.367	.442	.223
Absolute		<AbsD	-	>WB	>JD	>Man	>Lin	>Lid	>JM	>Lap
(Abs)	0.0315/-1.3%	.558		.646	.926	.869	.384	.371	.452	.217
Witten-Bell		<AbsD	<Abs	-	>JD	>Man	>Lin	>Lid	>JM	**>Lap**
(WB)	0.0311/-2.5%	.704	.646		.681	.621	.058	.066	.177	.027
Jensen-Shannon		<AbsD	<Abs	<WB	-	>Man	**>Lin**	**>Lid**	>JM	**>Lap**
(JD)	0.0286/-10.3%	.916	.926	.681		.274	.025	.014	.125	.033
Manhattan		<AbsD	<Abs	<WB	<JD	-	>Lin	>Lid	>JM	>Lap
(Man)	0.0284/-11.0%	.847	.869	.621	.274		.213	.455	.253	.319
Linear		<AbsD	<Abs	<WB	**<JD**	<Man	-	>Lid	>JM	>Lap
(Lin)	0.0264/-17.2%	.404	.384	.058	.025	.213		.762	.319	.249
Lidstone		<AbsD	<Abs	<WB	**<JD**	<Man	<Lin	-	>JM	**>Lap**
(Lid)	0.0250/-21.6%	.367	.371	.066	.014	.455	.762		.129	.040
Jelinek-Mercer		<AbsD	<Abs	<WB	<JD	<Man	<Lin	<Lid	-	>Lap
(JM)	0.0222/-30.4%	.442	.452	.177	.125	.253	.319	.129		.319
Laplace		<AbsD	<Abs	**<WB**	**<JD**	<Man	<Lin	**<Lid**	<JM	-
(Lap)	0.0210/-34.2%	.223	.217	.027	.033	.319	.249	.040	.319	

Table 71: Comparison of the *TextVis-CombWtScore* text and visual fusion method with the text-only results and the other fusion methods *TextVis-CombJointPr* and *TextVis-CombWtRank* for combining the *Shot+Adj+Video* interpolated text language model's results with the combined visual examples results of the different retrieval models for the TRECVid 2002 collection.

TRECVid 2002	*TextVis-CombWtScore*					*V. Text-Only*	*V. TextVis-CombJointPr*	*V. TextVis-CombWtRank*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ~ Wilc.*	*Impr. ~ Wilc.*	*Impr. ~ Wilc.*
Laplace	0.70, 0.30	**.1618**	.260	.187	.122	+0.8% ~ .654	+2,390.9% ~ .000	+34.3% ~ .011
Best	(0.80, 0.20	**.1716**	.272	.177	.122)	+2.3% ~ .319	+2,542.5% ~ .000	+2.3% ~ .319
Lidstone	0.70, 0.30	**.1616**	.256	.185	.121	+0.7% ~ .665	+2,338.5% ~ .000	+40.1% ~ .007
Best	(0.80, 0.20	**.1715**	.272	.177	.121)	+2.3% ~ .494	+2,323.9% ~ .000	+2.3% ~ .494
Linear	0.70, 0.30	**.1610**	.256	.187	.122	+0.3% ~ .729	+2,421.0% ~ .000	+38.5% ~ .033
Best	(0.80, 0.20	**.1713**	.272	.175	.122)	+2.1% ~ .494	+2,425.0% ~ .000	+2.1% ~ .494
Absolute Discounting	0.70, 0.30	**.1568**	.264	.185	.123	-2.3% ~ .233	+2,654.1% ~ .000	+34.2% ~ .049
Best	(0.90, 0.10	**.1704**	.272	.176	.121)	+1.6% ~ .295	+2,831.6% ~ .000	+1.6% ~ .295
Witten-Bell	0.70, 0.30	**.1590**	.260	.185	.124	-1.0% ~ .623	+2,451.8% ~ .000	+21.7% ~ .044
Best	(0.80, 0.20	**.1715**	.264	.176	.122)	+2.2% ~ .519	+2,652.3% ~ .000	+2.2% ~ .519
Jelinek-Mercer	0.70, 0.30	**.1663**	.264	.187	.121	+3.6% ~ .271	+2,247.8% ~ .000	+8.2% ~ .019
Best	(0.75, 0.25	**.1720**	.264	.177	.121)	+2.6% ~ .646	+1,250.1% ~ .000	+2.6% ~ .646
Absolute	0.70, 0.30	**.1570**	.260	.187	.124	-2.2% ~ .292	+2,532.6% ~ .000	+22.9% ~ .087
Best	(0.80, 0.20	**.1713**	.264	.177	.122)	+2.2% ~ .558	+2,722.5% ~ .000	+2.2% ~ .558
Manhattan	0.70, 0.30	**.1637**	.276	.185	.122	+1.9% ~ .584	N/A	+28.7% ~ .010
Best	(0.90, 0.10	**.1712**	.276	.177	.122)	+2.1% ~ .213	N/A	+2.1% ~ .213
Jensen Shannon	0.70, 0.30	**.1633**	.268	.187	.122	+1.7% ~ .464	N/A	+39.7% ~ .014
Best	(0.90, 0.10	**.1712**	.276	.177	.122)	+2.1% ~ .138	N/A	+2.1% ~ .138
Average of unbiased		**.1612**	.263	.186	.122	+0.4%	+2,433.8%	+29.8%
Average of best		**.1713**	.270	.177	.121	+2.2%	+2,392.5%	+2.2%

Table 72: Comparison of the *TextVis-CombWtScore* text and visual fusion method with the text-only results and the other fusion methods *TextVis-CombJointPr* and *TextVis-CombWtRank* for combining the *Shot+Adj+Video* interpolated text language model's results with the combined visual examples results of the different retrieval models for the TRECVid 2003 collection.

TRECVid 2003	*TextVis-CombWtScore*					*V. Text-Only*	*V. TextVis-CombJointPr*	*V. TextVis-CombWtRank*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*	*Impr. ∼ Wilc.*
Laplace	0.80, 0.20	**.1529**	.264	.191	.117	+8.8% ∼ .050	+295.9% ∼ .001	+8.8% ∼ .050
Best	(0.70, 0.30	**.1575**	.284	.199	.119)	+10.2% ∼ .173	+307.8% ∼ .001	-5.5% ∼ .511
Lidstone	0.80, 0.20	**.1536**	.268	.191	.116	+9.3% ∼ .036	+261.8% ∼ .002	+9.3% ∼ .036
Best	(0.70, 0.30	**.1591**	.280	.199	.120)	+11.4% ∼ .094	+300.5% ∼ .001	-4.2% ∼ .443
Linear	0.80, 0.20	**.1524**	.268	.189	.118	+8.4% ∼ .039	+281.1% ∼ .002	+8.4% ∼ .039
Best	(0.70, 0.30	**.1581**	.276	.200	.122)	+10.6% ∼ .139	+331.5% ∼ .000	-4.2% ∼ .355
Absolute Discounting	0.90, 0.10	**.1483**	.260	.184	.114	+5.5% ∼ .146	+379.9% ∼ .001	+5.5% ∼ .146
Best	(0.70, 0.30	**.1570**	.280	.193	.122)	+9.8% ∼ .262	+422.6% ∼ .000	-4.6% ∼ .081
Witten-Bell	0.80, 0.20	**.1534**	.268	.188	.117	+9.2% ∼ .158	+399.8% ∼ .001	+9.2% ∼ .158
Best	(0.70, 0.30	**.1611**	.280	.193	.122)	+12.7% ∼ .055	+424.7% ∼ .000	-3.7% ∼ .284
Jelinek-Mercer	0.75, 0.25	**.1564**	.288	.191	.121	+11.3% ∼ .081	+177.2% ∼ .005	+11.3% ∼ .081
Best	(0.70, 0.30	**.1605**	.276	.199	.122)	+12.3% ∼ .046	+335.6% ∼ .000	-6.6% ∼ .173
Absolute	0.80, 0.20	**.1533**	.268	.188	.117	+9.1% ∼ .158	+405.5% ∼ .001	+9.1% ∼ .158
Best	(0.70, 0.30	**.1592**	.280	.189	.122)	+11.4% ∼ .271	+427.8% ∼ .000	-2.3% ∼ .334
Manhattan	0.90, 0.10	**.1462**	.256	.180	.114	+4.1% ∼ .031	N/A	+4.1% ∼ .031
Best	(0.70, 0.30	**.1569**	.276	.193	.119)	+9.8% ∼ .095	N/A	-3.1% ∼ .256
Jensen Shannon	0.90, 0.10	**.1462**	.256	.179	.114	+4.0% ∼ .025	N/A	+4.0% ∼ .025
Best	(0.70, 0.30	**.1575**	.276	.199	.118)	+10.2% ∼ .095	N/A	-6.9% ∼ .212
Average of unbiased		**.1514**	.266	.187	.117	+7.7%	+314.5%	+7.7%
Average of best		**.1585**	.279	.196	.121	+10.9%	+364.4%	-4.5%

Table 73: Comparison of the *TextVis-CombWtScore* text and visual fusion method with the text-only results and the other fusion methods *TextVis-CombJointPr* and *TextVis-CombWtRank* for combining the *Shot+Adj+Video* interpolated text language model's results with the combined visual examples results of the different retrieval models for the TRECVid 2004 collection.

TRECVid 2004	*TextVis-CombWtScore*					*V. Text-Only*	*V. TextVis-CombJointPr*	*V. TextVis-CombWtRank*
Retrieval Method	*Prm*	*MAP*	*P10*	*P30*	*P100*	*Impr. ~ Wilc.*	*Impr. ~ Wilc.*	*Impr. ~ Wilc.*
Laplace	0.70, 0.30	**.0729**	.217	.155	.097	+6.3% ~ .584	+1,662.8% ~ .000	+0.4% ~ .303
Best	(0.55, 0.45	**.0790**	.248	.171	.094)	+15.2% ~ .758	N/A	+8.3% ~ .827
Lidstone	0.70, 0.30	**.0745**	.217	.157	.099	+8.7% ~ .524	+1,726.5% ~ .000	+0.5% ~ .324
Best	(0.55, 0.45	**.0812**	.265	.180	.095)	+18.3% ~ .767	N/A	+5.6% ~ .794
Linear	0.70, 0.30	**.0743**	.217	.159	.099	+8.3% ~ .572	+1,765.1% ~ .000	+1.5% ~ .536
Best	(0.55, 0.45	**.0810**	.257	.175	.093)	+18.1% ~ .767	N/A	+5.4% ~ .785
Absolute Discounting	0.70, 0.30	**.0763**	.217	.157	.100	+11.2% ~ .631	+2,342.8% ~ .000	+4.0% ~ .313
Best	(0.55, 0.45	**.0867**	.265	.187	.098)	+26.3% ~ .863	N/A	+8.2% ~ .738
Witten-Bell	0.70, 0.30	**.0775**	.226	.158	.102	+13.1% ~ .189	+2,421.0% ~ .000	+2.1% ~ .335
Best	(0.55, 0.45	**.0888**	.283	.190	.099)	+29.5% ~ .687	N/A	+14.9% ~ .282
Jelinek-Mercer	0.70, 0.30	**.0741**	.217	.152	.099	+8.1% ~ .381	+1,987.8% ~ .000	+3.5% ~ .262
Best	(0.60, 0.40	**.0750**	.230	.162	.093)	+9.3% ~ .919	N/A	+5.0% ~ .914
Absolute	0.70, 0.30	**.0762**	.222	.159	.102	+11.1% ~ .654	+2,476.0% ~ .000	+5.7% ~ .369
Best	(0.55, 0.45	**.0899**	.278	.193	.098)	+31.0% ~ .560	N/A	+12.7% ~ .476
Manhattan	0.70, 0.30	**.0756**	.213	.157	.100	+10.2% ~ .654	N/A	+4.9% ~ .392
Best	(0.55, 0.45	**.0834**	.239	.180	.098)	+21.6% ~ .767	N/A	+15.0% ~ .697
Jensen Shannon	0.70, 0.30	**.0758**	.222	.155	.100	+10.5% ~ .488	N/A	+5.5% ~ .197
Best	(0.55, 0.45	**.0844**	.261	.184	.097)	+23.1% ~ .665	N/A	+15.6% ~ .262
Average of unbiased		**.0752**	.219	.157	.100	+9.7%	+2,054.6%	+3.1%
Average of best		**.0833**	.258	.180	.096	+21.4%		+10.1%

Table 74: Comparison of unbiased retrieval models for the TRECVid 2002 search tasks using the *TextVis-CombWtScore* fusion method for combining text and visual results.

Ret. Method	*MAP/%Dif*	*Wilcoxon Test Results*								
Jelinek-Mercer		-	>Man	>JD	**>Lap**	>Lid	**>Lin**	>WB	**>Abs**	**>AbsD**
(JM)	0.1663		.294	.229	.024	.052	.011	.180	.020	.003
Manhattan		<JM	-	>JD	**>Lap**	**>Lid**	**>Lin**	>WB	**>Abs**	**>AbsD**
(Man)	0.1637/-1.6%	.294		.823	.031	.025	.014	.110	.014	.003
Jensen-Shannon		<JM	<Man	-	**>Lap**	**>Lid**	**>Lin**	>WB	**>Abs**	**>AbsD**
(JD)	0.1633/-1.8%	.229	.823		.021	.009	.004	.204	.008	.004
Laplace		**<JM**	**<Man**	**<JD**	-	>Lid	**>Lin**	>WB	>Abs	**>AbsD**
(Lap)	0.1618/-2.7%	.024	.031	.021		.693	.021	.706	.072	.036
Lidstone		<JM	**<Man**	**<JD**	<Lap	-	**>Lin**	>WB	>Abs	>AbsD
(Lid)	0.1616/-2.8%	.052	.025	.009	.693		.012	.744	.053	.057
Linear		**<JM**	**<Man**	**<JD**	**<Lap**	**<Lid**	-	>WB	>Abs	>AbsD
(Lin)	0.1610/-3.2%	.011	.014	.004	.021	.012		.890	.303	.130
Witten-Bell		<JM	<Man	<JD	<Lap	<Lid	<Lin	-	**>Abs**	**>AbsD**
(WB)	0.1590/-4.4%	.180	.110	.204	.706	.744	.890		.004	.001
Absolute		**<JM**	**<Man**	**<JD**	<Lap	<Lid	<Lin	**<WB**	-	>AbsD
(Abs)	0.1570/-5.6%	.020	.014	.008	.072	.053	.303	.004		.243
Absolute Discounting		**<JM**	**<Man**	**<JD**	**<Lap**	<Lid	<Lin	**<WB**	<Abs	-
(AbsD)	0.1568/-5.7%	.003	.003	.004	.036	.057	.130	.001	.243	

Table 75: Comparison of unbiased retrieval models for the TRECVid 2003 search tasks using the *TextVis-CombWtScore* fusion method for combining text and visual results.

Ret. Method	*MAP/%Dif*	*Wilcoxon Test Results*								
Jelinek-Mercer		-	>Lid	>WB	**>Abs**	>Lap	**>Lin**	**>AbsD**	>Man	>JD
(JM)	0.1564		.708	.432	.421	.557	.443	.410	.399	.366
Lidstone		<JM	-	**>WB**	**>Abs**	**>Lap**	**>Lin**	**>AbsD**	>Man	>JD
(Lid)	0.1536/-1.8%	.708		.036	.039	.033	.015	.044	.095	.076
Witten-Bell		<JM	**<Lid**	-	**>Abs**	>Lap	**>Lin**	**>AbsD**	**>Man**	**>JD**
(WB)	0.1534/-1.9%	.432	.036		.028	.936	.966	.392	.548	.572
Absolute		<JM	**<Lid**	**<WB**	-	>Lap	**>Lin**	**>AbsD**	**>Man**	**>JD**
(Abs)	0.1533/-2.0%	.421	.039	.028		.940	.961	.416	.548	.596
Laplace		<JM	**<Lid**	<WB	<Abs	-	**>Lin**	**>AbsD**	**>Man**	**>JD**
(Lap)	0.1529/-2.2%	.557	.033	.936	.940		.758	.138	.292	.292
Linear		<JM	**<Lid**	<WB	<Abs	<Lap	-	**>AbsD**	**>Man**	**>JD**
(Lin)	0.1524/-2.6%	.443	.015	.966	.961	.758		.072	.252	.346
Absolute Discounting		<JM	**<Lid**	<WB	<Abs	<Lap	<Lin	-	**>Man**	**>JD**
(AbsD)	0.1483/-5.2%	.410	.044	.392	.416	.138	.072		.928	.959
Manhattan		<JM	<Lid	<WB	<Abs	<Lap	<Lin	<AbsD	-	>JD
(Man)	0.1462/-6.5%	.399	.095	.548	.548	.292	.252	.928		.384
Jensen-Shannon		<JM	<Lid	<WB	<Abs	<Lap	<Lin	<AbsD	<Man	-
(JD)	0.1462/-6.5%	.366	.076	.572	.596	.292	.346	.959	.384	

Table 76: Comparison of unbiased retrieval models for the TRECVid 2004 search tasks using the *TextVis-CombWtScore* fusion method for combining text and visual results.

Ret. Method	*MAP/%Dif*	*Wilcoxon Test Results*								
Witten-Bell		-	>AbsD	>Abs	>JD	>Man	>Lid	>Lin	>JM	>Lap
(WB)	0.0775		.022	.003	.101	.081	.003	.003	.044	.001
Absolute Discounting		<WB	-	>Abs	>JD	>Man	>Lid	>Lin	>JM	>Lap
(AbsD)	0.0763/-1.5%	.022		.313	.619	.572	.282	.064	.282	.206
Absolute		<WB	<AbsD	-	>JD	>Man	>Lid	>Lin	>JM	>Lap
(Abs)	0.0762/-1.7%	.003	.313		.767	.619	.124	.101	.631	.151
Jensen-Shannon		<WB	<AbsD	<Abs	-	>Man	>Lid	>Lin	>JM	>Lap
(JD)	0.0758/-2.2%	.101	.619	.767		.358	.173	.124	.428	.012
Manhattan		<WB	<AbsD	<Abs	<JD	-	>Lid	>Lin	>JM	>Lap
(Man)	0.0756/-2.5%	.081	.572	.619	.358		.608	.224	.882	.335
Lidstone		<WB	<AbsD	<Abs	<JD	<Man	-	>Lin	>JM	>Lap
(Lid)	0.0745/-3.9%	.003	.282	.124	.173	.608		.081	.631	.101
Linear		<WB	<AbsD	<Abs	<JD	<Man	<Lid	-	>JM	>Lap
(Lin)	0.0743/-4.1%	.003	.064	.101	.124	.224	.081		.776	.440
Jelinek-Mercer		<WB	<AbsD	<Abs	<JD	<Man	<Lid	<Lin	-	>Lap
(JM)	0.0741/-4.4%	.044	.282	.631	.428	.882	.631	.776		.009
Laplace		<WB	<AbsD	<Abs	<JD	<Man	<Lid	<Lin	<JM	-
(Lap)	0.0729/-5.9%	.001	.206	.151	.012	.335	.101	.440	.009	

APPENDIX E

PAPERS PUBLISHED ON THIS WORK

Cooke, E , Furguson, P , Gaughan, G , Gurrın, C , Jones, G , Lee, H , Marlow, S , Mc Donald, K , McHugh, M , Murphy, N , O'Connor, N , O'Hare, N , Rothwell, S , Smeaton, A and Wılkens, P (2005), TRECVID 2004 experiments ın Dublın Cıty Unıversıty, ın 'Proceedings of TRECVıd 2004', NIST Specıal Publıcatıons

Mc Donald, K and Smeaton, A F (2005), A comparison of score, rank and probabılity-based fusıon methods for vıdeo shot retrıeval, ın 'Proceedings of the Fourth Internatıonal Conference on Image and Vıdeo Retrıeval (CIVR-2005)'

References

Adams, B , Amir, A , Dorai, C , Ghosal, S , Iyengar, G , Jaimes, A , Lang, C , Lin, C -Y , Natsev, A , Naphade, M , Chalapathy, N , Nock, H , Permuter, H , Singh, R , Smith, J , Srinivasan, S , Tseng, B , Ashwin, T and Zhang, D (2003), IBM Research TREC-2002 video retrieval system, *in* 'Proceedings of the Eleventh Text REtrieval Conference (TREC-2002)'

Aggarwal, C , Hinneburg, A and Keim, D (2001), On the surprising behavior of distance metrics in high dimensional space, *in* 'Lecture Notes in Computer Science', Vol 1973, Springer, pp 420–434

Akutsu, A and Tonomura, Y (1994), Video tomography An efficient method for camerawork extraction and motion analysis, *in* Proceedings of the second ACM international conference on Multimedia (MULTIMEDIA '94)', pp 349–356

Amir, A , Argillander, J O , Berg, M , Chang, S -F , Hsu, W , Iyengar, G , Kender, J , Lin, C -Y , Naphade, M , Natsev, A , Smith, J , Tesic, J , Wu, G , Yan, R and Zhang, D (2005), IBM Research TRECVID-2004 video retrieval system, *in* 'Proceedings of TRECVid 2004', NIST Special Publications

Amir, A , Berg, M , Chang, S -F , Hsu, W , Iyengar, G , Lin, C -Y , Naphade, M , Natsev, A , Neti, C , Nock, H , Smith, J , Tseng, B , Wu, Y and Zhang, D (2004), IBM Research TRECVID-2003 video retrieval system, *in* 'Proceedings of TRECVid 2003'

Armitage, L H and Enser, P G B (1996), Information need in the visual document domain Report on project rdd/g/235 to the British library research and innovation centre , Technical report, School of Information Management, University of Brighton

Ashwin, T V , Gupta, R and Ghosal, S (2002), Leveraging non-relevant images to enhance image retrieval performance, *in* 'Proceedings of the tenth ACM international conference on Multimedia (MULTIMEDIA '02)', pp 331–334

Baan, J , Van Ballegooij, A , Geusenbroek, J M , Den Hartog, J , Hiemstra, D , List, J , Patras, I , Raaijmakers, S , Snoek, C , Todoran, L , Vendrig, J , de Vries, A , Westerveld, T and Worring, M (2002), Lazy users and automatic video retrieval tools in (the) lowlands, *in* 'Proceedings of the Tenth Text REtrieval Conference (TREC-2001)', NIST Special Publications

Bach, J R , Fuller, C , Gupta, A , Hampapur, A , Horowitz, B , Humphrey, R , Jain, R C and Shu, C (1996), Virage image search engine An open framework for image management, *in* 'Symposium on Electronic Imaging Science and Technology - Storage & Retrieval for Image and Video Databases IV', IS&T/SPIE, pp 76–87

Beauchemin, S S and Barron, J L (1995), 'The computation of optical flow', *ACM Computing Surveys* **27**(3), 433–467

Berger, A and Lafferty, J (1999*a*), Information retrieval as statistical translation, *in* 'Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)', pp 222–229

Berger, A and Lafferty, J (1999*b*), The Weaver system for document retrieval, *in* 'Proceedings of the Eight Text REtrieval Conference (TREC-1999)', NIST Special Publications

Blum, A and Mitchell, T (1998), Combining labeled and unlabeled data with co-training, *in* 'Proceedings of the Workshop on Computational Learning Theory (COLT'98)', pp 92–100

Bookstein, A and Swanson, D (1974), 'Probabilistic models for automatic indexing', *Journal of the American Society for Information Science* **25**(5), 312–318

Bouthemy, P and Fable, R (1998), 'Motion characterization from temporal cooccurrences of local motion-based measures for video indexing', *Proceedings of the 12th International Conference on Pattern Recognition (ICPR-1998)* **1**, 905–908

Brandt, S , Laaksonen, J and Oja, E (2002), 'Statistical shape features in content-based image retrieval', *Journal of Mathematical Imaging and Vision* **17**(2), 187–198

Brodatz, P (1966), *Textures A Photographic Album for Artists and Designers*, Dover, New York

Brown, M G , Foote, J T , Jones, G J , Sparck Jones, K and Young, S J (1995), Automatic content-based retrieval of broadcast news, *in* 'Proceedings of the third ACM international conference on Multimedia (MULTIMEDIA '95)', ACM Press, pp 35–43

Canny, J (1986), 'A computational approach to edge detection', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **8**(6), 679–698

Chang, S -F , Chen, W , Meng, H J , Sundaram, H and Zhong, D (1997), VideoQ An automated content based video search system using visual cues, *in* 'Proceedings of the fifth ACM international conference on Multimedia (MULTIMEDIA '97)', ACM Press, pp 313–324

Chellappa, R , Wilson, C L and Sirohey, S (1995), 'Human and machine recognition of faces A survey', *Proceedings of IEEE* **83**(5), 705–740

Chen, M -Y and Hauptmann, A (2004), Searching for a specific person in broadcast news video, *in* 'International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)', Montreal, Canada

Chen, W and Chang, S -F (2000), Motion trajectory matching of video objects, *in* 'SPIE Storage and Retrieval for Media Databases 2000', Vol 3972, pp 544–553

Cheng, Y -J and Chen, H -H (2005), Aligning words from speech recognition and shots for video information retrieval, *in* 'Proceedings of TRECVid 2004', NIST Special Publications

Chua, T -S , Neo, S -Y , Li, K -Y , Wang, G , Shi, R , Zhao, M and Xu, H (2005), TRECVID 2004 search and feature extraction task by NUS PRIS, *in* 'Proceedings of TRECVid 2004', NIST Special Publications

Clarke, C , Craswell, N and Soboroff, I (2005), Overview of the TREC 2004 Terabyte track, *in* 'Proceedings of the Thirteenth Text REtrieval Conference (TREC-2004)', NIST Special Publications

Clarkson, K (1997), Nearest neighbor queries in metric spaces , *in* 'ACM Symposium on the Theory of Computing', pp 609–617

Cleverdon, C (1967), 'The Cranfield tests on index language devices', *Aslib Proceedings* pp 173–192

Common Evaluation Measures (2003), *in* E M Voorhees and L P Buckland, eds, 'Proceedings of the Eleventh Text REtrieval Conference (TREC-2002)', NIST Special Publications, pp A14–A21

Cooke, E , Furguson, P , Gaughan, G , Gurrin, C , Jones, G , Lee, H , Marlow, S , Mc Donald, K , McHugh, M , Murphy, N , O'Connor, N , O'Hare, N , Rothwell, S , Smeaton, A and Wilkens, P (2005), TRECVID 2004 experiments in Dublin City University, *in* 'Proceedings of TRECVid 2004', NIST Special Publications

Cooper, W (1995), 'Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval', *ACM Transactions on Information Systems* **13**(1), 100–111

Cover, T M and Thomas, J A (1991), *Elements of Information Theory*, Wiley, New York

Cox, I J , Miller, M L , Minka, T P , Papathomas, T V and Yianilos, P N (2000), 'The Bayesian image retrieval system, PicHunter Theory, implementation and psychophysical experiments', *IEEE Transactions on Image Processing*

Crestani, F , Lalmas, M , van Rijsbergen, C J and Campbell, I (1998), 'Is this document relevant? probably A survey of probabilistic models in information retrieval', *ACM Computing Surveys* **30**(4), 528–552

Croft, W B and Lafferty, J , eds (2003), *Language Modeling for Information Retrieval*, Kluwer Academic Publishers, Dordrecht, The Netherlands

Czirjek, C , O'Connor, N , Marlow, S and Murphy, N (2003), Face detection and clustering for video indexing applications, *in* 'Proceedings of Advanced Concepts for Intelligent Vision Systems (ACIVS 2003)'

Darwish, K , Doermann, D , Jones, R , Oard, D and Rautiainen, M (2002), TREC-10 experiments at University of Maryland CLIR and video, *in* 'Proceedings of the Tenth Text REtrieval Conference (TREC-2001)', NIST Special Publications

de Vries, A P and Westerveld, T (2004), A comparison of continuous vs discrete image models for probabilistic image and video retrieval, *in* 'Proceedings of IEEE International Conference on Image Processing (ICIP-2004)'

Dempster, A P , Laird, N M and Rubin, D B (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society* **B 39**(1), 1–38

Deselaers, T (2003), Features for image retrieval, Diploma thesis, RWTH Aachen University of Technology

Endres, D M and Schindelin, J E (2003), 'A new metric for probability distributions', *IEEE Transactions on Information Theory* **49**(7), 1858–1860

Equitz, W (1993), Retrieving images from a database using texture – algorithms from the QBIC system, Technical report, IBM Almaden Research Center

Faloutsos, C , Equitz, W , Flickner, M , Niblack, W , Petkovic, D and Barber, R (1994), 'Efficient and effective querying by image content', *Journal of Intelligent Information Systems* pp 231–262

Fisher, R A (1922), 'On the mathematical foundations of theoretical statistics', *Philosophical Transactions of the Royal Society* **222**, 309–368

Flickner, M , Sawhney, H , Niblack, W , Ashley, J , Huan, Q , Dom, B , Gorkani, M , Hafner, J , Lee, D , Petkovic, D , Steele, D and Yanker, P (1995), 'Query by image and video content The QBIC system', *IEEE Computer* **28**(9), 23–32

Fox, E and Shaw, J (1994), Combination of multiple searches, *in* 'Proceedings of the Third Text REtrieval Conference (TREC-1994)', NIST Special Publications, pp 243–252

Gale, W A and Church, K W (1994), What's wrong with adding one?, *in* N Oostdijk and P de Haan, eds, 'Corpus-Based Research into Language in honour of Jan Aarts', Rodopi, Amsterdam

Gauvain, J L , Lamel, L and Adda, G (2002*a*), 'The LIMSI broadcast news transcription system', *Speech Communication* **37**(1-2), 89–108

Gauvain, J , Lamel, L and Adda , G (2002*b*), 'The LIMSI broadcast news transcription system', *Speech Communication* **37**(1-2), 89–108

Giacinto, G and Roli, F (2004), Nearest-prototype relevance feedback for content based image retrieval, *in* 'Proceedings of the 18th International Conference on Pattern Recognition (ICPR-2004)', pp 989–992

Good, I J (1953), 'The population frequencies of species and the estimation of population parameters', *Biometrika* **40**, 237–264

Hafner, J , Sawhney, H , Equitz, W , Flickner, M and Niblack, W (1995), 'Efficient color histogram indexing for quadratic form distance functions', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17**(7), 729–736

Haralick, R M , Shanmugam, K and Dinstein, I (1973), 'Textural featues for image classification', *IEEE Transactions on Systems, Man and Cybernetics* **3**, 610–621

Hardy, G (1889), 'Insurance record', Correspondence Reprinted in Transactions of the Faculty of Actuaries, 8, 1920

Harter, S (1975), 'A probabilistic approach to automatic keyword indexing', *Journal of the American Society for Information Science* **26**, 197–206 and 280–289

Hauptmann, A , Baron, R , Chen, M -Y , Christel, M , Duygulu, P , Huang, C , Jin, R , Lin, W -H , Ng, T , Moraveji, N , Papernick, N , Snoek, C , Tzanetakis, G , Yang, J , Yang, R and Wactlar, H (2004), Informedia at TRECVID 2003 Analyzing and searching broadcast news video, *in* 'Proceedings of TRECVid 2003', NIST Special Publications

Hauptmann, A , Chen, M -Y , Christel, M , Huang, C , Lin, W -H , Ng, T , Papernick, N , Velivelli, A , Yang, J , Yan, R , Yang, H and Wactlar, H D (2005), Confounded expectations Informedia at TRECVID 2004, *in* 'Proceedings of TRECVid 2004', NIST Special Publications

Hauptmann, A G , Jin, R and Ng, T D (2002), Video retrieval using speech and image information, *in* 'Proceedings of the Tenth Text REtrieval Conference (TREC-2001)', NIST Special Publications

Hauptmann, A G , Yan, R , Qi, Y , Jin, R , Christel, M , Derthick, M , Chen, M -Y , Baron, R , Lin, W -H and Ng, T D (2003), Video classification and retrieval with the Informedia digital video library system, *in* 'Proceedings of the Eleventh Text REtrieval Conference (TREC-2002)', NIST Special Publications

Heesh, D , Howarth, P , Magalhães, J , May, A , Pickering, M , Yavlinsky, A and Ruger, S (2005), Video retrieval using search and browsing, *in* 'Proceedings of TRECVid 2004', NIST Special Publications

Hiemstra, D (1998), A linguistically motivated probabilistic model of information retrieval, *in* 'Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries (ECDL)', pp 569–584

Hiemstra, D (2002), Term-specific smoothing for the language modeling approach to information retrieval The importance of a query term, *in* 'Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)', ACM, pp 35–41

Hiemstra, D and Kraaij, W (1999), Twenty-one at TREC 7 Ad-hoc and cross-language track, *in* 'Proceedings of the Seventh Text REtrieval Conference (TREC-1998)', NIST Special Publications, pp 174–185

Howarth, P and Ruger, S (2004), Evaluation of texture features for content-based image retrieval, *in* 'Proceedings of the Third International Conference on Image and Video Retrieval (CIVR-2004)', pp 326–334

Howarth, P and Ruger, S (2005), Fractional distance measures for content-based image retrieval, *in* 'Advances in Information Retrieval, 27th European Conference on IR Research (ECIR 2005)', pp 447–456

Huang, J , Kumar, S R , Mitra, M , Zhu, W -J and Zabih, R (1997), Image indexing using color correlograms, *in* 'Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR-1997)', pp 762–768

Hull, D (1993), Using statistical testing in the evaluation of retrieval experiments, *in* 'Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '93)', ACM Press, pp 329–338

Ianeva, T , Boldareva, L , Westerveld, T , Cornacchia, R , Hiemstra, D and de Vries, A P (2005), Probabilistic approaches to video retrieval The Lowlands team at TREC VID 2004, *in* 'Proceedings of TRECVid 2004', NIST Special Publications

Ianeva, T I , de Vries, A P and Westerveld, T (2004), A dynamic probabilistic multimedia retrieval model, *in* 'Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2004)', pp 1607–1610

Ioka, M (1989), A method of defining the similarity of images on the basis of color information, Technical Report RT-0030, IBM Tokyo Research Lab

Irani, M , Anandan, P , Bergen, J , Kumar, R and Hsu, S (1996), 'Efficient representations of video sequences and their applications', *Signal Processing Image Communication* **8**, 327–351

Ishikawa, Y , Subramanya, R and Faloutsos, C (1998), MindReader Query databases through multiple examples, *in* 'Proceedings of the 24th Conference on Very Large Databases (VLDB)', pp 218–227

Jain, A K and Vailaya, A (1998), 'Shape-based retrieval a case study with trademark image databases', *Pattern Recognition* **31**(9), 1369–1390

Jelinek, F (1998), *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, Massachusetts

Jin, R and Hauptmann, A G (2002), Using a probabilistic source model for comparing images, *in* 'Proceedings of IEEE International Conference on Image Processing (ICIP-2002)'

Kirminen, P and Gabbouj, M (2000), The visual goodness evaluation of color-based retrieval processes, *in* 'Proceedings of the European Signal Processing Conference (EUSIPCO-2000)', pp 2153–2156

Kraaij, W , Smeaton, A F and Over, P (2004), TRECVid 2004 - an introduction, *in* 'Proceedings of TRECVid 2004', NIST Special Publications

Kraaij, W and Spitters, M (2003), Language models for topic tracking, *in* 'Language Modeling for Information Retrieval', Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 95–124

Lafferty, J and Zhai, C (2001), Document language models, query models, and risk minimization for information retrieval, *in* 'Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)', pp 111–119

Lafferty, J and Zhai, C (2003), Probabilistic relevance models based on document and query generation, *in* 'Language Modeling for Information Retrieval', Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 1–5

Laplace, P S (1995), *Philosophical Essay on Probabilities*, Spring Verlag, New York

Lavrenko, V (2000), Localized smoothing for multinomial language models, Technical Report CIIR IR-222, University of Massachusetts at Amherst

Lavrenko, V and Croft, W B (2001), Relevance-based language models, *in* 'Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)', pp 120–127

Lavrenko, V and Croft, W B (2003), Relevance models in information retrieval, *in* 'Language Modeling for Information Retrieval', Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 11–55

Lee, J H (1997), Analyses of multiple evidence combination, *in* 'Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '97)', pp 267–276

Lidstone, G (1920), 'Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities', *Transactions of the Faculty of Actuaries* **8**, 182–192

Liu, J (1991), 'Divergence measures based on the Shannon entropy', *IEEE Transactions on Information Theory* **37**(1), 145–151

Lorente, L and Torres, L (1999), Face recognition of video sequences in a MPEG-7 context using a global Eigen approach, *in* 'Proceedings of IEEE International Conference on Image Processing (ICIP-99)', pp 187–191

Ma, W-Y and Manjunath, B S (1997), 'NeTra A toolbox for navigating large image databases', *Proceedings of IEEE International Conference on Image Processing (ICIP-97)*

Ma, W Y and Zhang, H J (1998), Benchmarking of image features for content-based retrieval, *in* 'Proceedings of 32nd Asilomar Conference on Signals, Systems and Computers', pp 253–257

Mallat, S (1991), 'Zero-crossings of a wavelet transform', *IEEE Transactions on Information Theory* **37**(4), 1019–1033

Manjunath, B and Ohm, J-S (2001), 'Color and texture descriptors', *IEEE Transactions on circuits and systems for video technology* **11**, 703–715

Manmatha, R , Feng, F and Rath, T (2001), Using models of score distributions in information retrieval, *in* 'Proceedings of the LM Workshop 2001', pp 91–96

Manning, C D and Schutze, H (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, Massachusetts

Maron, M E and Kuhns, J L (1960), 'On relevance, probabilistic indexing and information retrieval', **7**, 216–244

Mc Donald, K and Smeaton, A F (2005), A comparison of score, rank and probability-based fusion methods for video shot retrieval, *in* 'Proceedings of the Fourth International Conference on Image and Video Retrieval (CIVR-2005)'

McKenna, S , Gong, S and Raja, Y (1997), Face recognition in dynamic scenes, *in* 'British Machine Vision Conference', pp 140–151

Miller, D R , Leek, T and Schwartz, R M (1999*a*), BBN at TREC-7 Using hidden Markov models for information retrieval, *in* 'Proceedings of the Seventh Text REtrieval Conference (TREC-1998)', NIST Special Publications, pp 113–142

Miller, D R , Leek, T and Schwartz, R M (1999*b*), A hidden Markov model information retrieval system, *in* 'Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)', pp 214–221

MPEG7 Committee (2002), *MPEG-7 Multimedia Content Description Interface*, ISO

Naphade, M R and Smith, J R (2004), On the detection of semantic concepts at TRECVID, *in* 'Proceedings of the 12th annual ACM international conference on Multimedia (MULTIMEDIA '04)', New York, NY, pp 660–666

Nelson, R and Polana, R (1992), 'Qualitative recognition of motion using temporal texture', *CVGIP Image Understanding* **56**(1), 78–89

Ney, H , Essen, U and Kneser, R (1994), 'On structuring probabilistic dependencies in stochastic language modeling', *Computer Speech and Language* **8**, 1–38

Nigam, D and Ghani, R (2000), Understanding the behaviour of co-training, *in* 'Proceedings of KDD-2000 Workshop on Text Mining'

O'Callaghan, R J and Bull, D R (2002), Improved illumination-invariant descriptors for robust colour object recognition, *in* 'Proceedings of International Conference on Acoustics, Speech and Signal Processing'

Ojala, T , Aittola, M and Matinmikko, E (2002), Empirical evaluation of MPEG-7 XM color descriptors in content-based retrieval of semantic image categories, *in* 'Proceedings of the 16th International Conference on Pattern Recognition (ICPR-2002)', Vol 2, pp 1021–1024

Ojala, T , Pietikainen, M and Harwood, D (1996), 'A comparative study of texture measures with classification based feature distributions ', *Pattern Recognition* **29**(1), 51–59

Ortega, M , Rui, Y , Chakrabarti, K , Mehrotra, S and Huang, T S (1997), Supporting similarity queries in MARS, *in* 'Proceedings of the fifth ACM international conference on Multimedia (MULTIMEDIA '97)', pp 403–413

Pass, G , Zabih, R and Miller, J (1996), Comparing images using color coherence vectors, *in* 'Proceedings of the fourth ACM international conference on Multimedia (MULTIMEDIA '96)', pp 65–73

Pentland, A , Picard, R W and Sclaroff, S (1996), 'Photobook Content-based manipulation of image databases', *International Journal of Computer Vision* pp 233–254

Pickering, M , Heesh, D , O'Callaghan, R , Ruger, S and Bull, D (2003), Video retrieval using global features in keyframes, *in* 'Proceedings of the Eleventh Text REtrieval Conference (TREC-2002)', NIST Special Publications

Ponte, J M (1998), A Language Modeling Approach to Information Retrieval, PhD thesis, University of Massachusetts at Amherst

Ponte, J M and Croft, W B (1998), A language modeling approach to information retrieval, *in* 'Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)', pp 275–281

Porter, M F (1980), 'An algorithm for suffix stripping', *Program* **14**, 130–137

Puzicha, J , Hofmann, T and Buhmann, J (1997), Non-parametric similarity measures for unsupervised texture segmentation and image retrieval, *in* 'Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR-1997)', pp 267–272

Puzicha, J , Rubner, Y , Tomasi, C and Buhmann , J (1999), Empirical evaluation of dissimilarity measures for color and texture, *in* 'Proceedings of the International Conference on Computer Vision (ICCV-1999)', Vol 2, Corfu, Greece, pp 1165–1173

Quenot, G M , Moraru, D , Ayache, S , Charhad, M , Guironnet, M , Carminati, L , Mulhem, P , Gensel, J , Pellerin, D and Besacier, L (2005), CLIPS-LIS-LSR-LABRI experiments at TRECVID 2004, *in* 'Proceedings of TRECVid 2004', NIST Special Publications

Rao, C R (1982), 'Diversity Its measurement, decomposition, apportionment and analysis', *Sankyha The Indian Journal of Statistics* **44**(A), 1–22

Ratcliff, J and Metzener, D (1988), 'Pattern matching The gestalt approach', *Dr Dobb's Journal* p 46

Rautiainen, M and Doermann, D (2002), Temporal color correlograms for video retrieval, *in* 'Proceedings of the 16th International Conference on Pattern Recognition (ICPR-2002)', Quebec, Canada

Rautiamen, M , Hosio, M , Hanski, I , Varanka, M , Kortelainen, J , Ojala, T and Seppanen, T (2005), TRECVID 2004 experiments at team Oulu, *in* 'Proceedings of TRECVid 2004', NIST Special Publications

Rautiainen, M , Ojala, T and Seppanen, T (2004), Analysing the performance of visual, concept and text features in content-based video retrieval, *in* 'Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval', New York, NY, pp 197–204

Rautiainen, M , Penttila, J , Pietarila, P , Noponen, K , Hosio, M , Koskela, T , Makela, S -M , Peltola, J , Liu, J , Ojala, T and Seppanen, T (2004), TRECVID 2003 experiments at MediaTeam Oulu and VTT, *in* Proceedings of TRECVid 2003', NIST Special Publications

Rautiainen, M , Penttila, J , Vorobiev, D , Noponen, K , Vayrynen, P , Hosio, M , Matinmikko, E , Makela, S -M , Peltola, J , Ojala, T and Seppanen, T (2003), TREC 2002 video track experiments at MediaTeam Oulu and VTT, *in* 'Proceedings of the Eleventh Text REtrieval Conference (TREC-2002)', NIST Special Publications

Rautiainen, M , Seppanen, T , Penttila, J and Peltola, J (2003), Detecting semantic concepts from video using temporal gradients and audio classification, *in* 'International Conference on Image and Video Retrieval', Urbana, IL

Robertson, S (1977), 'The probability ranking principle in IR', *Journal of Documentation* **33**(4), 294–304

Robertson, S and Sparck Jones, K (1976), 'Relevance weighting of search terms', *Journal of the American Society for Information Science* **27**, 129–146

Robertson, S , Walker, S , Jones, S , Hancock-Beaulieu, M M and Gatford, M (1995), Okapi at TREC-3, *in* 'Proceedings of the Third Text REtrieval Conference (TREC-1994)', NIST Special Publications

Rocchio, J (1971), Relevance feedback in information retrieval, *in* 'The SMART Retrieval System Experiments in Automatic Document Processing', Prentice Hall Inc

Rubner, Y , Puzicha, J , Tomasi, C and Buhmann , J (2001), 'Empirical evaluation of dissimilarity measures for color and texture ', *Computer Vision and Image Understanding* **84**, 25–43

Rubner, Y , Tomasi, C and Guibas, L J (2000), 'The Earth Mover's distance as a metric for image retrieval', *International Journal on Computer Vision* **40**(2), 99–121

Rui, Y , Huang, T and Chang, S -F (1997), 'Image retrieval past, present, and future', *International Symposium on Multimedia Information Processing*

Rui, Y , Huang, T S and Mehrotra, S (1997), Content-based image retrieval with relevance feedback in MARS, *in* 'Proceedings of IEEE International Conference on Image Processing (ICIP-97)', pp 815–818

Salton, G (1971), *The SMART Retrieval System - Experiments in Automatic Document Processing*, Prentice Hall Inc , Englewood Cliffs, NJ

Salton, G and Buckley, C (1988), 'Term-weighting approaches in automatic text retrieval', *Information Processing & Management* **24**(5), 513–523

Salton, G , Wong, A and Yang, C (1975), 'A vector space model for automatic indexing', *Communications of the ACM* **18**(11), 613–620

Shio, A and Sklansky, J (1991), Segmenting people in motion, *in* 'Proceedings of IEEE Workshop on Visual Motion', pp 325–332

Smeaton, A F (1998), Independence of contributing retrieval strategies in data fusion for effective information retrieval, *in* 'Proceedings of the 20th BCS-IRSG Colloquium on IR', Springer-Verlag Workshops in Computing, Grenoble, France

Smeaton, A F , Kraaij, W and Over, P (2004*a*), The TREC VIDeo retrieval evaluation (TRECVID) A case study and status report, *in* 'Proceedings of RIAO 2004'

Smeaton, A F , Kraaij, W and Over, P (2004*b*), TRECVid - an introduction, *in* 'Proceedings of TRECVid 2003', NIST Special Publications

Smeaton, A F and Over, P (2003), The TREC-2002 video track report, *in* E M Voorhees and L P Buckland, eds, 'Proceedings of the Eleventh Text REtrieval Conference (TREC-2002)', NIST Special Publications, pp 69–85

Smeaton, A F , Over, P and Taban, R (2002), The TREC-2001 video track report, *in* E M Voorhees and D K Harman, eds, 'Proceedings of the Tenth Text REtrieval Conference (TREC-2001)', NIST Special Publications

Smith, J and Chang, S -F (1997), 'Visually searching the web for content', *IEEE Multimedia* **4**(3), 12–20

Smith, J R and Chang, S -F (1996*a*), 'Querying by color regions using the VisualSEEk content-based visual query system'

Smith, J R and Chang, S -F (1996*b*), VisualSEEk A fully automated content-based image query system, *in* 'Proceedings of the fourth ACM international conference on Multimedia (MULTIMEDIA '96)'

Smith, J R , Srinivasan, S , Amir, A , Basu, S , Iyengar, G , Lin, C -Y , Naphade, M , Ponceleon, D and Tseng, B (2002), Integrating features, models, and semantics for trec video retrieval, *in* 'Proceedings of the Tenth Text REtrieval Conference (TREC-2001)', NIST Special Publications

Smith, M A and Kanade, T (1995), Video skimming for quick browsing based on audio and image characterization, Technical Report CMU-CS-95-186, Carnegie Mellon University

Snoek, C , Worring, M , Geusebroek, J , Koelma, D and Seinstra, F (2005), The MediaMill TRECVID 2004 semantic video search engine, *in* 'Proceedings of TRECVid 2004', NIST Special Publications

Song, F and Croft, W B (1999), A general language model for information retrieval, *in* 'Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)', pp 279–280

Sonka, M , Hlavac, V and Boyle, R (1998), *Image Processing, Analysis and Machine Vision*, 2nd edn, PWS Publishing

Sparck Jones, K , Robertson, S , Hiemstra, D and Zaragoza, H (2002), Language modeling and relevance, *in* 'Language Modeling for Information Retrieval', Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 57–71

Stegmann, M B and Gomez, D D (2002), A brief introduction to statistical shape analysis, Technical report, Informatics and Mathematical Modelling, Technical University of Denmark

Swain, M J and Ballard, D H (1991), 'Colour indexing', *International Journal on Computer Vision* **7**(1), 11–32

Tamura, H , Mori, S and Yamawaki, T (1978), 'Textural features corresponding to visual perception', *IEEE Transactions on Systems, Man and Cybernetics* **8**(6), 460–472

Torres, L and Vila, J (2001), 'Automatic face recognition for video indexing applications', *Pattern Recognition* **35**(3), 615–625

Turk, M and Pentland, A (1991), 'Eigenfaces for recognition', *Journal of Cognitive Neuroscience* **3**(1), 71–86

van Rijsbergen, C J (1979), *Information Retrieval*, second edn, Butterworths, London

Vasconcelos, N (2000), Bayesian Models for Visual Information Retrieval, PhD thesis, Massachusetts Institute of Technology

Vasconcelos, N and Lippman, A (1998), Embedded mixture modeling for efficient probabilistic content-based indexing and retrieval, *in* 'Proceedings of SPIE Multimedia Storage and Archiving Systems III', Vol 3527, pp 134–143

Vasconcelos, N and Lippman, A (2000), Bayesian representations and learning mechanisms for content-based image retrieval, *in* 'SPIE'

Wactlar, H D (2000), Informedia - search and summarization in the video medium, *in* 'Proceedings of Imagina 2000 Conference'

Wan, X and Kuo, C -C J (1996), Color distribution analysis and quantization for image retrieval, *in* 'Proceedings of SPIE Storage and Retrieval for Still Image and Video Databases IV', Vol 2670, pp 8–16

Westerveld, T (2004), Using generative probabilistic models for multimedia retrieval, PhD thesis, University of Twente

Westerveld, T and de Vries, A P (2004), Multimedia retrieval using multiple examples, *in* Proceedings of the Third International Conference on Image and Video Retrieval (CIVR-2004)', Springer-Verlag, pp 344–352

Westerveld, T , de Vries, A P and van Ballegooij, A (2003), CWI at the TREC-2002 video track, *in* 'Proceedings of the Eleventh Text REtrieval Conference (TREC-2002)', NIST Special Publications, pp 207–216

Westerveld, T , de Vries, A P , Van Ballegooij, A , de Jong, F and Hiemstra, D (2003), A probabilistic multimedia retrieval model and its evaluation, *in* 'EURASIP Journal on Applied Signal Processing', Vol 2, pp 186–198

Westerveld, T , Ianeva, T , Boldareva, L , de Vries, A P and Hiemstra, D (2004), Combining information sources for video retrieval The Lowlands team at TRECVID 2003, *in* 'Proceedings of TRECVid 2003'

Wiskott, L , Fellous, J -M , Kruger, N and von der Malsburg, C (1997), 'Face recognition by elastic bunch graph matching', *IEEE Tranactions on Pattern Analysis and Machine Intelligence* **19**(7), 775–779

Witten, I H and Bell, T C (1991), 'The zero-frequency problem Estimating the probabilities of novel events in adaptive text compression', *IEEE Transactions on Information Theory* **37**, 1085–1094

Won, C S , Park, D K and Park, S -J (2002), 'Efficient use of MPEG-7 edge histogram descriptor', *ETRI Journal* **24**(1), 23–30

Worring, M , Nguyen, G P , Hollink, L , Van Gemert, J and Koelma, D (2004), Interactive search using indexing, filtering, browsing, and ranking, *in* 'Proceedings of the Twelfth Text REtrieval Conference (TREC-2003)', NIST Special Publications

Wu, L , Guo, Y , Qui, X , Feng, Z , Rong, J , Jin, W , Zhou, D , Wang, R and Jin, M (2004), Fudan University at TRECVID 2003, *in* 'Proceedings of TRECVid 2003', NIST Special Publications

Xu, J , Weischedel, R and Nguyen, C (2001), Evaluating a probabilistic model for cross-lingual information retrieval, *in* 'Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)', pp 105–110

Yan, R and Hauptmann, A G (2003), The combination limit in multimedia retrieval, *in* 'Proceedings of the eleventh ACM international conference on Multimedia (MULTIMEDIA '03)', Berkeley, CA

Yan, R and Hauptmann, A G (2004), Co-Retrieval A boosted reranking approach for video retrieval, *in* 'Proceedings of the Third International Conference on Image and Video Retrieval (CIVR-2004)', Springer-Verlag, pp 60–69

Yan, R , Hauptmann, A G and Jin, R (2003), Negative pseudo-relevance feedback in content-based video retrieval, *in* 'Proceedings of the eleventh ACM international conference on Multimedia (MULTIMEDIA '03)', Berkeley, CA

Yan, R , Yang, J and Hauptmann, A G (2004), Learning query-class dependent weights in automatic video retrieval, *in* 'Proceedings of the 12th annual ACM international conference on Multimedia (MULTIMEDIA '04)', New York, NY, pp 548–555

Yang, M -H , Kriegman, D J and Ahuja, N (2002), 'Detecting faces in images A survey', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(1), 34–58

Yavlinsky, A , Pickering, M J , Heesch, D and Ruger, S (2004), A comparative study of evidence combination strategies, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing'

Zaragoza, H , Hiemstra, D , Tipping, M and Robertson, S (2003), Bayesian extension to the language model for ad hoc information retrieval, *in* 'Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)'

Zhai, C and Lafferty, J (2001*a*), Model-based feedback in the language modeling approach to information retrieval, *in* 'Tenth International Conference on Information and Knowledge Management (CIKM 2001)', pp 403–410

Zhai, C and Lafferty, J (2001*b*), A study of smoothing methods for language models applied to ad hoc information retrieval, *in* 'Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)', ACM Press, pp 334–342

Zhai, C and Lafferty, J (2002), Two-stage language models for information retrieval, *in* 'Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)', pp 49–56

Zhang, H J , Low, C Y , Smoliar, S W and Wu, J H (1995), Video parsing, retrieval and browsing An integrated and content-based solution, *in* 'Proceedings of the third ACM international conference on Multimedia (MULTIMEDIA '95)'

Zhang, H J , Zhong, D , Smoliar, S W and Gong, Y (1997), 'An integrated system for content-based video retrieval and browsing', *Pattern Recognition* **30**(4), 643–658

Zhou, X S and Huang, T S (2003), 'Relevance feedback in image retrieval A comprehensive review', *ACM Multimedia Systems* **8**(6), 536–544