Novelty Detection in Video Retrieval: Finding New News in TV News Stories.

Georgina Gaughan BSc.

A dissertation submitted in partial fulfilment of the requirements for the award of
Doctor of Philosophy

to the

Center For Digital Video Processing
School of Computing
Dublin City University

Supervisor: Prof. Alan F. Smeaton

August, 2006.

This thesis is based on the candidate's own work, and has not previously been submitted for a degree at any academic institution.

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

Georgina Gaughan 5/9/06.

Georgina Gaughan

Abstract

Novelty detection is defined as the detection of documents that provide "new" or previously unseen information. "New information" in a search result list is defined as the incremental information found in a document based on what the user has already learned from reviewing previous documents in a given ranked list of documents. It is assumed that, as a user views a list of documents, their information need changes or evolves, and their state of knowledge increases as they gain new information from the documents they see. The automatic detection of "novelty", or newness, as part of an information retrieval system could greatly improve a searcher's experience by presenting "documents" in order of how much extra information they add to what is already known, instead of how similar they are to a user's query. This could be particularly useful in applications such as the search of broadcast news and automatic summary generation.

There are many different aspects of information management, however, this thesis, presents research into the area of novelty detection within the content based video domain. It explores the benefits of integrating the many multi modal resources associated with video content those of low level feature detection evidences such as colour and edge, automatic concepts detections such as face, commercials, and anchor person, automatic speech recognition transcripts and manually annotated MPEG7 concepts into a novelty detection model. The effectiveness of this novelty detection model is evaluated on a collection of TV new data.

Acknowledgements

Achieving a PhD degree is a very rewarding yet often trying experience. It was made so much more bearable by the wonderful people around me and as a result these people deserve a special mention. They were able to provide me with the support, strength and light hearted relief required to carry on in many occasions. First and foremost, heartfelt thanks to my parents, Michael and Josephine and my brothers and sisters, Philomena, Colm, Sinéad, Declan and Stephan who were always there for me.

I wish to sincerely thank my supervisor Prof. Alan Smeaton for all his support and guidance from the first day of applying for the *IRCSET* scholarship from Australia right though to the end.

I would like to thank the old gang Carol, Ciara, Laura, Isobella, Julie and Dimitri who have always been there no matter what, listened to me and offered valuable support and suggestions throughout.

The PhD experience would not have been as enjoyable without the marvelous people I met and become close friends with, within the CDVP group over the last number of years. In particular, the following people deserve a special acknowledgement, for their invaluable contributions and suggestions, reading and rereading of the thesis, dinners, jokes, constant support and often times good natured nagging. In no particular order, the old gang including Cathal, Tom, Kieran, Hyowon, Paul and Jiamin. A special thanks to Cathal and Tom for their advice and invaluable suggestions, reading and rereading of the thesis. The new gang, once again in no particular order Sinead, Sandra, James, Pete, Mike, Paul, Neil, Colum, Mary, Niall and Fabrice. A special thanks to Sandra, Sinead, Pete and James for keeping me sane and reading the thesis. Thank you all so much.

I would like to thank Rong Yan of Carnegie Mellon University for allowing me to use the combination weights for high level feature combinations.

I gratefully acknowledge The Irish Research Council for Science, Engineering and Technology (IRCSET). Without its scholarship this PhD would not have been possible. I would also like to acknowledge the support of Enterprise Ireland.

Contents

List of Tables

List of Figures

Chapter 1

Introduction to IR

This thesis introduces a novelty detection model for content based video retrieval. This chapter firstly provides a high level background to general information retrieval. It will then firstly introduce and secondly provide the motivation behind the concept of novelty detection within information retrieval. It will outline the objectives of the research carried out in this thesis and finally will describe the organisation of this thesis.

1.1 Introduction

Today's society has become so familiar with the notion of information retrieval that many of its keywords and concepts have been gradually integrated into our commonly used vocabulary. Requiring a piece of information on a given topic will sometimes draw a response "Google it", a familiar concept to most people today.

> "Information retrieval is the name of the process or method whereby a prospective user of the information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him"

This is one of the oldest definitions of information retrieval written by Mooers in the 1950's and cited in Savino and Sebastiani 1998 [SS98]. With the development of many commercial information retrieval systems or search engines like Yahoo! and Google in the 1990's, information retrieval has become well known to the majority of the population. Using the World Wide Web has become more accessible and useable, thanks to the facilities of these search engines. We are now able to navigate and browse through more than eight billion webpages using links alone.

An information retrieval system is an implementation of a software algorithm that gathers, indexes, searches and manages a document collection, text, video, image or audio, be they static (TREC, medical, government, education libraries) or dynamic (www, digital video libraries) in nature [Ago02]. The system is designed with the overall aim of aiding potential users in the retrieval of information they require from the collection of data. It does not however, answer a particular question, it simply provides information on the existence and location of documents that the user should find satisfies their information need. These documents can then be considered relevant to the information need.

1.2 Basic Components of an Information Retrieval System

The four main processes/components of an Information system are:

- Input (documents): The offline task of the conversion of documents into formal representations, which can be manipulated easily by the computer is called the "Indexing Process". Documents are partially stored as a list of words and the frequency of those word occurrences, in these documents
- Input (Queries): the task of representing the user's information need as a formal representation using a similar algorithm/technique that was ap-

plied to the documents during the off line indexing process, is termed creating the "Query".

- Processing: This is the task or matching process, where the system calculates the relevance- similarity of the query to the formal representations or index of all the documents in the collection. Documents are matched to a query if their similarity is above a predetermined threshold.

- Output: This task- process ranks the retrieved documents selected by the processor in order of decreasing degree of relevance, according to their predetermined scores to the user's query and displays these retrieved documents to the user, in what is called a "Ranked List".

- Feedback-Relevance: Feedback is a component of a query and involves a human judgement and the processor in the retrieval system shown in Figure 1.1.

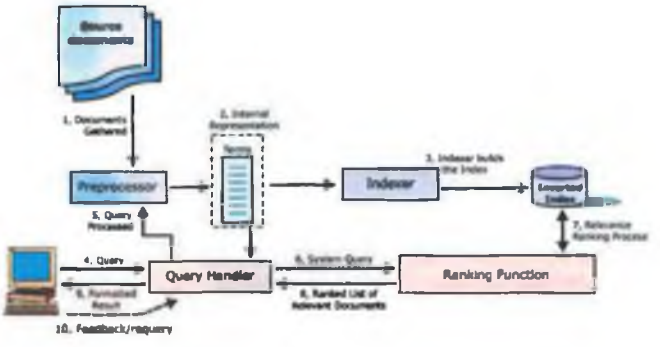

Figure 1.1: A Basic Information Retrieval System

Figure 1.1 outlines the components and steps involved in a basic information retrieval system. It is generally considered that, there are four main steps that must be carried out in any information system [Ago02].

- Gathering: The core of the information retrieval system is the corpus or the data collection from which it must search and retrieve documents

of interest according to the queries the users submit. A corpus can be either virtual or dynamic. In virtual corpora, documents are discarded after indexing. Most medical journals and business corporations are static archives leading to good examples of a virtual corpora. The most commonly used corpora are those of the dynamic World Wide Web. This corpus must be collected or "gathered" from the WWW software tools or web search agents called "spiders" or "crawlers" where the documents are identified, located and downloaded. The documents are then preprocessed to remove frequently occurring words, stopwords, or unnecessary duplicates and create either:

1. A virtual collection: - this occurs when the documents are discarded after indexing.
2. A physical collection: - documents are maintained.
3. Or both

- Indexing: This phase applies many processes to the newly gathered collection such as stopword removal, stemming (the reduction of commonly used words to their root) for example "computer" stems to "comput", lexical analysis (changing all capital letters to lowercase), content analysis and term weighting, to create a formal representation of documents which are stored in a data structure specifically chosen to enhance the speed at which they are referenced/retrieved. Each document indexed is usually given a unique document identifier.

- Searching: This phase is the user interaction stage. There are three steps in the searching phase:

1. Takes the user's query and processes it in a similar manner to the document indexing phase, with algorithms such as lexical analysis, stopword removal, stemming and term weighting applied to the query.
2. Matches the query representation against the document representation using some similarity technique

3. It then returns the retrieved documents in decreasing rank order to the user.

- Document Management: A dynamic WWW collection must be continuously updated or maintained, due to its volatile nature, identified by [RNBY99] as one of the challenges of the WWW. Web servers are continuously been added or deleted to the web while web pages themselves are continuously being updated by their authors, renamed, relocated or sometimes deleted. This can cause severe problems for the end user as documents found to be relevant in the un-maintained and non-updated corpus may no longer exist or contain any relevant information for the required topic in the WWW location. A typical example of a maintained corpus is the Google cache. This corpus is up-dated and maintained regularly yet quite often a searcher is still returned locations of documents that are no longer in existence, hence the benefit of the cache.

We will now look at ways in which some of the information retrieval stages, can be formally represented using mathematical models.

1.3 Mathematical Models of Information Retrieval

Mathematic models have been used in information retrieval in an attempt to accurately recreate the real world concept of information retrieval, the matching of user needs and relevant documents. Both documents and queries are represented formally as mathematical models of the same type allowing matching functions to accurately access the similarity between document and query models. Mathematical models enable the implementation of information retrieval systems such as Google. There are many approaches to how documents, queries and retrieval can be modeled. These are broken up into standard IR categories as seen in Figure 1.2 and include :

- *Classical* which includes Boolean, Vector space and Probabilistic models.

- *Non-Classical* which includes Information Logic based on logical imaging, Situation Theory based on an information calculus approach for information retrieval and Interaction Information Retrieval, a quantum mechanics approach to information retrieval which involves, modeling the interaction of query with documents.

- *Alternative* which include Fuzzy, Cluster, Artificial Intelligence, Language Models and Latent Semantic Indexing models.

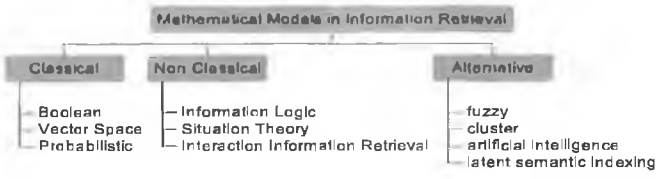

Figure 1.2: Categories of Models in Information Retrieval

1.3.1 The Boolean Model

The Boolean Model was the traditional and most widely used model in commercial information systems until the 90's. One of the first to be built representing the problem of structured queries, the Boolean model is based on both boolean logic and set theory. Both the document and the user query are represented as sets of terms but query terms are connected by logical operators (e.g logical AND, NOT, OR) to form a structured boolean formula. Up to recently professional searchers who knew the document corpus and the structure of the system acted as interpreters between the user and the system. Boolean retrieval involves the user entering a structured query. Take for example the query, *cat* AND *dog* AND *mouse* denoted by A, B and C respectively.

The documents returned in this case will contain all three indexed terms as shown in Figure 1.3. However a document that contains both *cat* and *dog*, yet does not contain the term *mouse* will not be returned as relevant and this

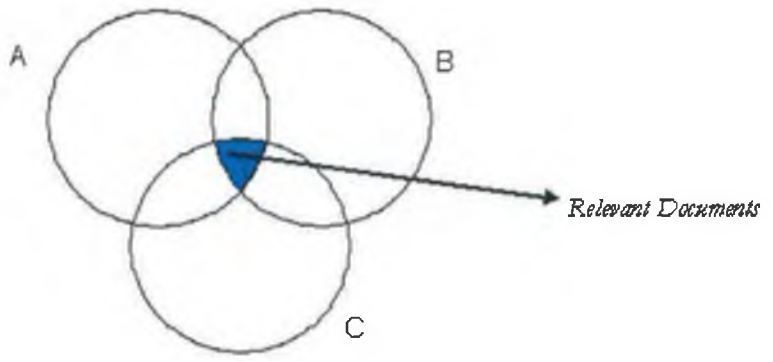

Figure 1.3: Boolean Model

can lead to frustration as the user may have increased his knowledge with the information contained in this "allegedly" irrelevant document.

Boolean querying is very unforgiving as there is no "nearly" relevant. This simple example illustrates the exact matching concept, and the notion of relevant/irrelevant, of the Boolean model. To the professional searcher this may be an advantage, however to a novice exact matching degrades the retrieval performance, as a document cannot be ranked according to its degree of probable relevancy. It either contains a term, many terms or it does not. Another disadvantage of this model is that the formulation of structured queries for multi-concept topics is rather complex and would require the use of professional searchers.

1.3.2 The Vector Space Model

An improvement to the Boolean model is the Vector Space Model with the use of "term weighting". This involves applying an importance value to each term in a document or collection according to its frequency/occurrence in that document and across all documents in a collection, which is a non-trivial operation. Documents containing words, that occur frequently are usually function words and offer no value or use to the user. These function words, e.g. "or", are called stopwords and are generally removed from the document. Documents

containing words that occur infrequently usually provide little information on documents content, however words occurring with a mid range frequency more or less describe the content of the document. This is the basic idea used in term weighting algorithms, such as *tf.idf* discuss later in the section.

The Vector Space Model [SWY75], another classical model of information retrieval, is so-called as it represents both documents and queries as mathematical vectors in a t-dimensional vector space (t is the number of terms in the collection). Retrieval is then based on how close a document vector $\vec{d}$ is to a query vector $\vec{q}$, see Figure 1.4. Documents are plotted in a space of index terms.

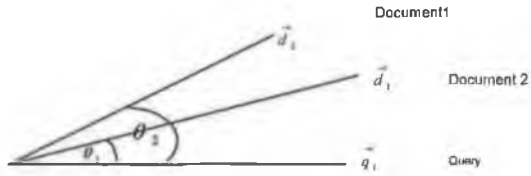

Figure 1.4: Vector Space Model

The documents are ranked according to the highest/closest score to the query vector. This similarity measure is calculated as the cosine of the angle θ between the query vector $\vec{q}$ and the doc vector $\vec{d}$. In the above diagram it is clear that d_1 is more similar to q_1 than d_2. These models assume all index terms(words) are equal which is not true and so they need an additional term weighting that is in most cases, *tf.idf*.

It is necessary to calculate two values for each term in an index in order to weight the terms appropriately.

- The term frequency or number of occurrences of a term in a particular document *tf*, and also
- The frequency or occurrences of the term over all documents, *df*.

The inverse of the document frequency *df*, *idf*, is used along with the term

frequency *tf*, to implement one of the best known term weighting algorithms, "tf.idf". The formal algorithm is calculated as follows:-

$$idf_j = \frac{log(N)}{df_j} \tag{1.1}$$

The logarithm of the number of documents in a collection N, divided by the number of documents where this term occurs (df_j)

$$weight_{ij} = tf_{ij} \times idf_j \tag{1.2}$$

$weight_{ij}$ represents the weights assigned to a term t_j in a document d_i.
tf_{ij} represents the frequency of term t_j in document d_i.
N represents the number of documents in a collection.
df_j represents the number of documents where term t_j occurs at least once.

This formula has had many modifications and extensions since it performed rather poorly due to its inability to normalise the length of a document and hence its consequences of favouring long documents over shorter ones. The newer versions have made improvements on this.

One major assumption of the vector space model is that query terms are considered independent of one another, however real world situations have dependent terms for example, informational and retrieval, software and engineering.

1.3.3 The Probabilistic Model

Robertson and Cooper [Rob77] introduced this classical information retrieval model, when they published the Probability Ranking Principal (PRP):

> "if the reference retrieval systems response to each request is a ranking of the documents in the collections in order of decreasing probability of usefulness to the user who submitted the request

> where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data "

Basically, retrieval of documents is based on the value of the probability of a document to be relevant to a particular query being above the value of the probability of a document being irrelevant to the same query. Once documents with the required probabilities are retrieved they are ranked in decreasing order of usefulness to the query. The documents retrieved are those that reach above a certain cut off point or threshold [Dom01]. A popular implementation of the probabilistic model is BM25[1] [RWB+97], which uses a different formula to index both documents and queries as follows.

To index a query:

$W_{(ij)}$ the weight assigned to a query term is given by :

$$W_{qj} = \frac{tf_{qk}}{k_3 + tf_{qk}} \cdot \ln[(N - df_j)/df_j] \tag{1.3}$$

tf_{qk} represents the frequency of term t_k in document d_q.
N represents the number of documents in a collection.
df_j represents the number of documents where term t_j occurs at least once.
k_3 is a parameter.

To index documents:

$W_{(ij)}$ the weight assigned to a term in a document is given by :

$$W_{ij} = \frac{(k_1 + 1)tf_{ij}}{K + tf_{ij}} \; where \; K = k_1[(1 - b) + b \cdot \frac{l_i}{avdl}] \tag{1.4}$$

[1]BM stands for Best Match

tf_{ij} indicates the frequency of a term j within a document i.
b and k_1 are parameters.
K is the ratio between the length l_i (sum of tf_{ij}) of document i and the collection mean, denoted by *avdl*.

BM25 has repeatedly been shown to be very effective and possibly the most effective term weighting formula in IR research.

1.4 IR Evaluation Measures

Users of an information retrieval system expect a set of relevant documents that accurately match their topic request of information need. In order to evaluate the effectiveness and efficiency of various retrieval systems various evaluation measures have been introduces.

Two standard evaluation measures, Recall and Precision, form the basis of most evaluation measures with information retrieval. Both measures are set based working over a non-ranked fixed set of documents. A user's preference on measuring the effectiveness of a system dictates which measure they are likely to concern themselves with, users requiring an extensive list of documents relevant to the query, at the expense of non material being returned, are likely to evaluate based on recall, whereas users who would like to receive only relevant documents within a return list are more likely to measure the system's performance on precision. Precision measures the proportion of retrieved documents that are relevant to the users query (seen in equation 1.5) while recall measures the proportion of relevant documents returned for a users query (seen in equation 1.6).

$$Precision = \frac{Relevant \cap Retrieved}{Relevant} \tag{1.5}$$

$$Recall = \frac{Relevant \cap Retrieved}{Retrieved} \tag{1.6}$$

Precision and Recall are commonly presented pair wise in the form of a precision recall graph in order to avoid the drawbacks that exist using the precision measure alone, returning a value of 1.0 precision when results set consists of a single document which happens to be relevant to the query and again using the recall measure alone, returning a value of 1.0 when a system returns the entire collection in response to a query. However it is difficult to accurately make comparison between effectiveness and accuracy using a precision and recall pair.

Average Precision A ranked based evaluation measure based on standard precision, that measures the effectiveness of a retrieval system in returning relevant documents high within a results set, to a particular topic. It is calculated by averaging the precision as each relevant document is found within a ranked list. Any relevant documents in the collection that are not retrieved in the ranked list, give a precision value of zero. The measure never decreases as more documents are added to the end of an existing list. So, consider there are five relevant documents in a collection to a specific topic. Three of them have been retrieved within the ranked list at ranks 1, 5, 6. Then the precision at rank 1 is P@1 $= 1(\frac{1}{1})$ at rank 5 P@5 $= 0.4(\frac{2}{5})$ at rank 6 P@6 $= 0.5(\frac{3}{7})$. The average precision of all documents retrieved by the system for this topic then becomes $(1 + 0.4 + 0.5 + 0 + 0)/5 = 0.38$. To accurately assess the performance of a retrieval system though, it is more effective to consider a retrieval system's ability in returning relevant documents to a set of topics rather than one in particular. This is calculated using mean average precision.

Mean Average Precision (MAP) is as the name suggests the mean of the average precision of all topics within an evaluation run over many topics.

1.5 Introduction to Novelty

In 1999 Hal Varian, an economist, suggested that from an economist's viewpoint *"the value of information is that it is only new information that matters"* [Var99]. The context of his statement was a challenge to the established tradition in information retrieval whereby documents are ranked in response to a query by their similarity to that query. This approach to document ranking is firmly established partly because it can be implemented in a computationally efficient manner which was important in the early days of information retrieval. Nowadays it remains prevalent because it allows search engines like Google to implement sub-second response time when searching billions of web pages for millions of users daily.

Yet despite its computational efficiency and scalability, ranking by query similarity is merely one tool which we use as part of our broader information seeking tasks in which we engage in many times daily. When we search we formulate a query in our mind, input some keywords into a search box, browse the resulting list of summaries, select a document or page and view it, maybe go back to our search ranking and view some more documents and in doing this, we may clarify our own information needs a bit more so we may reformulate our query and issue another search. This generates another document ranking which includes the documents we've seen and viewed before and don't want to see again! The search function, activated when we click the SEARCH box, is helping us because it is fast, but it is not intelligent and it still leaves us to do all the interpretation of search outputs. Over time we have grown tolerant to the fact that IR searching is actually a low-level function in the broader picture of information seeking.

Recent trends in IR reveal a more questioning approach to the established tradition and include developments like document summarisation, clustering of the outputs of search results and emergence of attempts to capture users' contexts in search. All these try to ease the cognitive load on searchers by

making the interpretation of search output more digestible. One other technique for doing this, which we are interested in, is the automatic detection of novelty in search output.

Novelty in search output is defined as the incremental information added to a document based on what the user has already learned from looking at previous documents in the document list. It assumes that we do not forget information (this assumption is re-addressed in Chapter 3) and that as a user views a results list of documents their information need changes or evolves, and their state of knowledge increases as they learn new things from the documents they see. At any point in the list, the technique of *relevance feedback* can be used to help reformulate the query to take account of shifting information needs, and this is commonplace in information retrieval. However, little work has been done on taking account of what the user has already seen from documents viewed, i.e. there is little work in the automatic detection of *novelty* in the documents being presented to users. It follows that if we use relevance feedback to account for shifting information needs we should use each document's novelty value as a factor in determining where it should appear in a document ranking.

Objectives of the Research being undertaken

A typical broadcast TV news program is usually a very rich source of information on a variety of diverse news topics. However it is also rife with repetition as video footage, story elements and developments in stories and even story introductions within the same broadcast are re-used. Once a user has viewed a relevant shot, any subsequent shots that provide no new information are made redundant and become useless to the user from the point of view of increasing his knowledge on a particular topic. These shots however take valuable visual space on the user's interface, hindering a user's interaction with the system as it must wade through these redundant shots in search of new and unseen information, which may be displayed way down the results list. Within the text domain novelty detection actively seeks data which provides new information

on a topic to the user.

In this thesis we focus on novelty detection within the video domain. Although many novelty models exists for the detection of new documents in the text domain, currently none of the novelty models developed account for novelty detection in the video domain. We propose adapting the novelty detection concept developed for the text domain to address novelty detection within the video information retrieval taking account of the many resources that exist within video. In novelty detection within video IR we seek to organise broadcast news retrieval results based on the degree of "newness" to the topic rather than the traditional ranking by degree of relevance, thereby increasing the user ability to make an informed decision on whether accessing a shot is useful.

Novelty detection within broadcast news obviously eliminates redundancy among shots but also enables the ability to track a story over several broadcasts from either the same broadcaster or across multiple broadcasters. It can be used to highlight the outcomes and conclusions rather than the earlier and outdated story elements of a complete story. Novelty detection modules are now contained within many automatic summarisation systems for the summarisation of a video or multiple videos. The Intelligence Community are looking at novelty detection modules to help decrease the assessor's work load of identifying shots that provide new information, for example, Helicopters landing, army tanks moving in background, increased numbers of people on the street and explosions. Currently assessors wade through hours of endless uneventful footage.

Thesis Research Questions

The following research question will be answered through the course of this thesis.

1. How to adapt the novelty detection concepts already carried out within the text domain to develop novelty detection models for the much more complex video domain ?

2. Can novel shots be automatically detected from within a list of shots within the video domain ?

3. Do models designed to detect novel shots from a chronologically ordered list of shots using text resources alone out-perform other resources and combinations of resources also available within the video domain or does novelty detection need to utilise the other resources available from within video to accurately complete the task ?

4. How do novelty detection models developed for the identification of novel shots from a chronologically ordered list of relevant shots for a topic within the video domain, perform compared to a human assessor's performance of the task ?

5. How do the performances of the many modalities available for each video sequence compare to each other in the task of detecting novel shots from a chronologically ordered list of relevant shots for a topic ?

1.6 Thesis Organisation

The remainder of this thesis is organised as follows. Chapter 2 will provide an overview of multimedia information retrieval and in particular video retrieval. It will briefly describe the history of digital video and its structure, before outlining the various components of a modern video retrieval system. The chapter will then describe TRECVid and its contribution to the video retrieval research arena before, finally examining the current state of the art in video retrieval systems within the video domain.

Chapter 3 introduces the concept of novelty detection in information retrieval and in particular novelty detection within the text domain. The chapter will then provide an overview of the TREC novelty track, which was developed to focus research and development into the detection of novel documents from a

results list and will then describe our novelty detection model developed for the novelty track in 2004.

Chapter 4 provides answers to question one. It shows there is a need for novelty detection in video retrieval and introduces novelty detection in the context of content based video retrieval. It investigates the challenges and considerations that must be observed during the design of a novelty detection model in the video domain. It will then discuss the detection of novel information from within a search output for any user specific topic. This chapter presents a novelty detection model designed to accurately identify novel shots from a results list within the video domain. Chapter 5 presents the experimental methodology for investigating the performance of our novelty detection approaches. Chapter 6 will present answers to the remaining research questions described above by reporting on the findings of the experiments carried out on each of the novelty detection models designed for each of the video resources available within the video domain. Finally Chapter 7 will conclude this thesis where we will summarise our approach to novelty detection in the video retrieval domain. We will reflect on the answers to the research questions identified above. We will then conclude by making suggestions for further work in the area of novelty detection within video retrieval.

Chapter 2

Introducing Multimedia IR

In this chapter we review the ever-growing popular research area of Multimedia Information Retrieval. We will define the concept of video information retrieval, describing briefly the history of digital video and its structure. We then discuss the various components used within modern video retrieval systems, to improve the systems performance in returning accurate results to a users query. We will discuss TRECVid and its contribution to the video retrieval research arena and finally, we will examine three state of the art video retrieval systems developed by Dublin City University, CMU and IBM respectively

2.1 Introduction to Multimedia Information Retrieval

Traditionally information retrieval operated over text documents from large collections, with state of the art commercial information retrieval systems successfully searching and answering a users information need. The development of new media technologies and integrated multiple media such as images, mp3's, audio and video have created vast multimedia libraries and archives in areas such as medical, criminal investigation, art galleries and TV broadcasting to name but a few. It is apparent that there is a need for information management, organisation, retrieval and navigation through these vast multimedia

archives. For example, the BBC archive stores an additional one million new items per year including video, image, audio and text information. This thesis concentrates on content-based retrieval of digital video. An information retrieval system that provides access to a video collection is far more complex than a traditional information retrieval system dealing with textual data alone. The main reason for this complexity lies in the inability to automatically analyze the video content accurately. The interpretation of video is more difficult due to the richness of its content including visual, audio, text and semantic information [Bim99]. Before delving into the aspects of the digital video retrieval, we first need to have some understanding of digital video itself.

2.1.1 An Introduction to Digital Video

During the 1970's consumers became familiar with the concept of video with the introduction of Video Home System (VHS) by JVC. Video is a sequence of twenty five to thirty images per second, giving an illusion of motion, synchronised with an audio track. Analogue video requires on average 1.3 MB of space for each image in uncompressed form. Sequential storage of this video format required a large capacity storage medium which was available only in the form of a magnetic tape at the time.

With the development of new technologies and compression standards (described in section 2.1.1), digital video became a reality and was introduced in the 1980's. It offered many interesting features over traditional analogue video, for example:

- Higher picture quality, easier storage and transmission across networks
- Retrieval of scenes/chapters instantaneously as it is stored on random access media such as CD-interactive disks or DVDs (Digital Versatile Disks).

Digital video manipulation has become so commonplace that it is easily produced and edited by not only video production companies but also home users.

It is widely used in major business corporations and by individuals through applications like video conferencing, video lectures, entertainment, documentaries, advertisements, e-mail attachments and so on. These applications of digital video coupled with the decrease in the costs of acquiring the software and hardware necessary to manipulate digital video, has led to the generation of large digital video libraries both in organizational and personal environments at enormous rates.

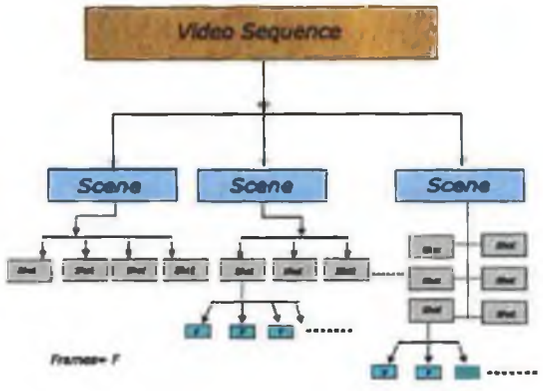

Figure 2.1: Structure of a digital video sequence

Figure 2.1 outlines the hierarchical structure of a digital video sequence. The following is a description of each of the components individually.

Shot

The basic unit of a video sequence is the frame. A number of frames, recorded in a single continuous action by a single camera at a specific time are collectively called a shot. Shots are commonly recognised within the information retrieval community as the basic unit of retrieval from a video collection.

Keyframe: A keyframe is a single representative frame for a group of frames composing a shot. The task of selecting which frame best represents the shot

is very subjective and the most commonly used technique involves selecting the middle frame from the group of frames. This however is not always very representative and in some cases where the shot is quite long, two or more keyframes are needed to accurately represent the content of the shot. Although keyframe representation is adequate for most video representation it is however unsuitable when analysing the temporal aspects of a video sequence.

Scene

A scene is a collection of shots grouped into a logical combination depicting a story/event or object of focus. Scenes can be classified as being static or dynamic, depending on the state of motion within the sequence of shots. From a human perspective the retrieval of scenes from video is far more attractive than shots as scenes provide a level of meaning and understanding of the developing video sequence. However it is very difficult to automatically detect scenes as the shots composing a scene can be visually quite different and thus scene boundary detection is largely unreliable across most genre (television news is the exception).

Video

In addition to a visual layer we also have an audio layer in video. These two layers are synchronized using the system layer. Lately we have seen the inclusion of a semantic information layer usually expressed in an XML-like standard known as MPEG-7 [Com02].

Digital Video Compression

The Moving Frame/Picture Expert Group established in 1988, provided an international video compression standard ISO/IEC 11172, commonly known as

MPEG-1 in 1993 [pag], which enabled the storage of digital video on CD- Interactive medium. MPEG-1 sets a typical image resolution of 352×288 pixels at 25 frames per second, resulting in VHS quality. In 1995, MPEG-2 was released providing a greater level of compression, which enabled digital video to be stored on DVD's and transmitted over many networks. Since then, the MPEG-4 standard has been released enabling object based encoding and supporting 2D and 3D video modes. Recently the MPEG-7 standard was released, which describes the multimedia content of digital video by adding an extendable and interoperable metadata layer to the digital video stream [Com02, YS03].

Within the MPEG-1 standard a frame is viewed as being of one of the following three types:

- **I-frame (Intra frame)**: This frame is treated as still image and is encoded with lossy compression using JPEG compression block by block completely independent of frames adjacent to it.
- **P-frame (Predicted frame)**: The frame is coded with reference to a previous P-frame or I-frame with motion estimation.
- **B-frame (Bi-directional frame)**: This frame is coded with reference to the preceding or next I or P-frame with motion estimation. The more B-frames included in an MPEG sequence the higher the level of compression.

I-frames are the most important frames in MPEG as they are the frame of least compression. They provide a reference point from which the motion-compensation is determined for the P and B frames. Depending on the compression required and the encoder used all three types of frames form a MPEG frame sequence called a GOP(Group of Pictures). MPEG standard states there should be at least one I-frame within a series of ten frames. An example of two valid MPEG frame sequences include

- I P P

- I B B P B B I

Within this thesis we work with I-frames as they are of better quality than P or B frames and in the following section we will discuss the accurate retrieval video data that answers a users information need.

2.2 Video Retrieval- the challenges

Information Retrieval within the text domain has made great advances within the last decade with the introduction of cheap capable technology and the web. Many solutions to the conventional method of ranking relevant documents according to the degree of relevance have been proposed to suit different user requirements and implemented successfully over many collections such as question answering, summarization and novelty detection. The main challenge within the video information retrieval research community is to achieve a similar standard of retrieval within the video domain to that which exists within the text domain. The retrieval of video data is much more complex than that of traditional text data for many reasons, including the fact that there are many more media components to be considered when manipulating the contents of a video document. Data is not only considered on a conceptual level, by issuing queries with keywords like *cat* or *computer* which is the standard form of retrieval for text documents. We are now also working on a perceptual level due to the visual nature of video by composing queries that also contain images, video clips or audio examples, which might contain the desired feature or object that the user wishes to retrieve from a video collection. For example a user wishing to retrieve one or more images or video clips of "Bill Clinton" may issue a query with one or more example images of Bill Clinton addressing a press conference with the US Flag in the background, which he/she may have found on the internet.

The following are a list of the major challenges to the development of video IR.

Processing Requirements: Twenty five to thirty frames a second are needed, each frame consisting of a typical resolution of 352×288 pixels, to give the visual illusion of motion in a video sequence in MPEG-1 format. This data component along with the various other components such as audio description, semantic descriptions and feature detection evidences associated with a similar sequence of video, means it is of orders of magnitude many times larger and more complex than its text domain equivalent, even with today's very high quality compression. This is a major problem as it hampered video IR research growth due to the inability to carry out experiments on large scale collections of data which has been possible within the text domain for years. Until recently the storage of a terabyte of data was considered a very expensive task, however with the technical advances and reduction in cost, the potential to create large video collections that mimic a real world scenario such as the BBC archive which comprises of over 1 million or 2145 terabytes of data, for the research and development of video IR related experiments is becoming less of a challenge.

User Interface: One of the goals of designing a user interface for any system should be the provision of simple, straightforward and easy interaction that does not confuse the user. However a multimedia system such as a video search system, requires sophisticated interface elements for searching and displaying of results [GSG+03]. Since video is visual information it is considered the norm to present the search results visually to the user and allow him/her to browse through the results [Sme00]. It is a video information retrieval challenge to provide the user with an interface, that allows him/her to quickly and effectively retrieve and browse through a set of results, similar to a user's interaction with a text retrieval system. The shot is the common unit of retrieval in video IR and is usually represented in an interface via a keyframe. This brings us to the question, over which frame of the shot should be displayed, observing the fact that a user may find a shot relevant/irrelevant depending on the selected keyframe. The issue of redundant shots being displayed on the users interface is yet another issue and it take up valuable space on a users results list. This issue raised in Chapter 4. Due to its complex nature there are many challenges

yet to be tackled when designing a video retrieval system interface to reduce the users effort and cognitive load during search and retrieval.

Feature Extraction: Text is considered the foundation for the retrieval of relevant shots from any content-based video collection. This conceptual level of accurate text presentation has received a lot of attention and although not a solved problem, has many mature solutions. The perceptual level however is not solved and is non trivial. Textual descriptions of perceptual events by human judgements is subjective and hence we cannot rely on our own descriptions of colour, texture and emotional content for example. In an attempt to more consistently interpret the perceptual view of a video sequence, various feature/concept detectors have been proposed. The area of concept detection has become a hot research topic in video research over the last decade however major challenges still remain namely accuracy and coverage. Currently the accuracy of feature/concept detectors is very low and in many cases the inclusion of such feature evidences will inevitably degrade the accuracy of a retrieval system. To date many concept detectors have been developed for specific domains such as the detection of a goal being scored in a soccer video, and a limited set of concepts have been developed by groups to detect a certain concept in video data such as the "beach", the "sky" or the presence of a named individual such as "Bill Clinton".

Automatic Semantic Extraction: Humans have an ability to easily decipher the semantic meaning of a visual image. It is far more difficult however for a machine to automatically extract the semantic meaning from an image or similarly a video sequence. Within the text domain the semantic meanings are integrated into the text and text IR preforms adequately at providing documents matching a user's request. Although we can extract some semantic meaning from audio tracks, automatic detectors have not reached a level that can adequately bridge the semantic gap between the low level features such as colour and texture and their higher level meanings such as "sunset" or "airplane taking off" within a visual image. In order to meet this challenge and overcome

the semantic gap while the feature detectors remain at their current level of accuracy, human annotation of video is currently being employed through the use of predefined manually created ontologys such as the "LSCOM" ontology which contains 1000 and 460 concepts respectively and the "MediaMill" ontology which contains 101 concepts. These ontologys are discussed further in Chapter 4. This does not solve the challenge of accurate and scalable video content representation however as human annotation is very expensive with regard to both time and manpower. More detailed information on annotation is described in Chapter 4.

Yet another challenge to the development of video IR involves advancing research into alternative applications in retrieval such as summarisation, question answering, finding named entities, story tracking and novelty detection as successfully researched within the text domain.
Summarisation systems provide a user with an accurate description of the contents of the document enabling him/her to selectively choose documents that are most likely to answer their specific need. An example of text summarisation include a summarisation system developed by Allan *et.al.* [AGK01] that monitors new programmes for any changes. Within the video domain, summarisation systems are limited to specific narrow domains, as technology has not reached the point where all features can be accurately detected for every possible event. [SOMM04] is an example of a video summarisation system over a narrow domain. It uses feature detectors which highlight the important events within soccer matches such as the scoring of a goal or a penalty being taken. Although this works well within the narrow domain of sports it is as yet not possible to accurately summarise the visual contents of a broad domain of video such as a news story collection.

Question-Answering systems have been widely researched and implemented within the text retrieval domain. They take as input a text query in the form of a question for example "who is Bill Clinton? " and return a ranked list of text fragments that are likely to answers the user's query. Currently there are

no Question-Answering system applications within the video domain.

Yet another application, that challenges the conventional approach of ranking relevant results by their degree of relevance to a topic, is that of intra-novelty detection researched within the text domain and this is described further in Chapter 3. However yet again, novelty detection has not been researched within the video retrieval domain.

2.3 Video Information Retrieval System Components

The last decade has seen many variations in the methods employed to search and retrieve video data. In this section we look at the video components that can be extracted and utilized in an attempt to improve a user's interaction experience with a video retrieval system. Initially video collections were very small and users of the system had a general idea what was contained within the collection. Manual annotation was employed on such collections whereby videos were named as descriptively as possible and manually grouped into categories according to their various relationships such as the genre they belonged to, which enabled the rapid retrieval of a specific video. On the selection of a specific video of interest the user could browse through the video shots to rapidly find a particular segment of interest for playback [CDV04]. Figure 2.2 shows an example of such a system. As collections grew, more sophisticated approaches for the retrieval of specific information were required, as this method soon became impractical, because users no longer knew what was in the collection and inevitably ended up blindly navigating through a collection, in search of relevant material. In the following sections we discuss three features which can be used in many modern video retrieval systems.

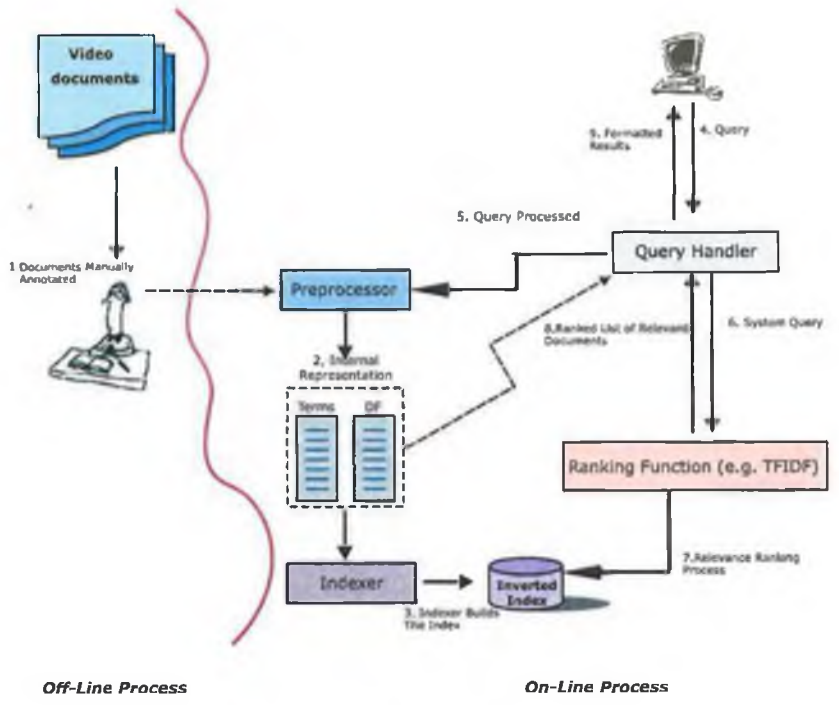

Figure 2.2: An example of an early video retrieval system

Text Searching

There are three different forms of text, all complementing each other, that can be utilised within video retrieval.

1. *Automatic Speech Recognition(ASR)* transcripts are derived automatically from the audio track of a video sequence. These transcripts are time stamped and aligned to each shot in the video collection during an offline process, providing the most in depth textual resource for video retrieval. There are two disadvantages when using ASR transcripts within video retrieval. Firstly although a transcripts for a specific video is aligned to each shot within the video, the words spoken may not match any of the visual features that occur within that shot and secondly automatic speech recognition detectors find it difficult to accurately spell named identities or locations.

2. *Closed caption(CC)* texts are written and transmitted during the broad-

cast of a television program as an aid to people with hearing disabilities. Once again time stamping allows the accurate alignment of text to each shot within the video collection, although in cases where the time stamp is missing, shot alignment is carried out based on the ASR transcripts [RM84][RHH+04] during an offline process. CC text is written by humans with the subsequent result that people, named entities or locations are usually spelled accurately. It is not written verbatim, according to the spoken audio which leads to alternative words being used to get the same point across. Both of these characteristics can aid the overall performance of a retrieval system. In addition, as CC text is designed for the hearing impaired, supplementary text may be added into the transcript in an attempt to aid the viewer by describing what is audibly happening within a scene such as the identification of a door knock.

3. *Optical character recognition(OCR)* another text resource available to the video domain is automatically extracted from an image. Consider a TV news interview, in the majority of cases the interviewee's name, location and current subject of discussion will appear on the bottom of the screen. OCR provides a valuable source of evidence as it is most likely describing what is occurring within a shot accurately which would greatly improve the performance of a retrieval system when searching for a specific topic.

Traditional text retrieval preprocessing techniques such as "stopword removal" and "stemming" are carried out on each of the three sources of text within video. Over the years text-based retrieval has consistently proved to be the single best performing component within TRECVID and more detailed information on text retrieval was described in Chapter 1. Retrieval models such as BM25, *tf*idf*, probabilistic and language models have all been employed for the accurate retrieval of video using text with BM25 achieving the best performance in TRECVID2004 and text once again achieving the best performance in TRECVID2005 [SKO04b, SKO05].

Although all the text resources are aligned to shots within the video it is not necessarily the case that words within the text accurately define the contents of a particular shot. As a result it is inadequate to rely on text descriptions alone for the effective retrieval of a relevant shot. Consider for example an interview on the war in Iraq; although Saddam's name may mentioned at the beginning of the interview, a shot of the leader may not necessarily be shown until the end of the news story. In this respect text fails to deliver adequate retrieval performance. In order to improve retrieval performance in video retrieval, text retrieval has been augmented with image based retrieval [GSG+03]. This enables a user to search a video collection with a particular image that may be similar to their information need, such as an image of "Bill Clinton" if the need for video data on Bill Clinton is required.

Image Searching

Typically within an image retrieval system, often referred to as low-level feature extraction, each shot in a video collection is analyzed by a number of predetermined low-level feature detectors such as colour, edge and texture. These detectors assign an evidence score to the shot based on a feature's existence or non-existence. Such a system takes an example image or video as input which closely matches a user's information request similar of the TRECVid topic description seen in Figure 2.3. The query image is analyzed in a similar manner

Figure 2.3: Topic 144: Find shots of Bill Clinton speaking with at least part of a US flag visible behind him. Three example images that matches the user's information need

to every other shot within the video collection using the same low-level feature

detectors. The similarity between the query image and each shot's keyframe within the collection is then analyzed based on the feature detection evidences. Shots with a similarity score over a certain threshold are deemed "probably relevant" and returned to the user in the form of a ranked list. Query images are provided as part of the TRECVid topics or as a keyframe which is the product of relevance feedback, requesting a "more like this" scenario which has been implemented successfully in video IR systems such as [BCG+03, BCG+04]. An example of querying by image can be seen in Figure 2.4. Retrieval performance within image retrieval is very dependant on the quality of the query image and whether or not the video collection has images with similar low-level features.

Figure 2.4: An example of querying a video retrieval system using image only

As stated previously, some common low-level features which have been detected successfully and aid overall video retrieval performance include colour, edge and texture.

Colour is the most popular and effective low level feature used within video retrieval systems. Visible light consists of a continuous spectrum of wavelengths which stimulates the retina of our eyes. These wavelengths range from approximately seven hundred and eighty nanometers for red to three hundred and fifty

nanometers for violet, see Figure 2.5. Objects absorb and reflect light waves differently and it is this variation of reflection, transmission and absorbtion which allows us to perceive an object's colour. There are three main type of colour perception that we are sensitive to "Hue", "Saturation" and "Brightness". Hue refers the dominant colour. Saturation refers to the purity of the colour and Brightness refers to the brightness or luminance of the surface [Ear85]. Computers display images using a fixed number of "pixels" which are square units used to store information about each individual colour. The greater the number of pixels that are used to compose an image the better the quality of that image. Colour spaces, including the *RGB* or *YCrCb*, have been designed in an attempt to express colour by accurately modeling a human's perception of colour [Poy]. The "RGB" colour space is designed for computer hardware and is often used to display colour on television screens and computer monitors. The "YCbCr" colour space is used for encoded videos. The most common colour features used within video retrieval include, dominant colour, scalable colour, colour structure and colour layout and GOF/GOP colour all of which are defined within the MPEG7 (XML-like standard) standard [Com02].

Figure 2.5: Colour wavelength spectrum

Large colour or texture difference areas create boundaries or edges humans can perceive as objects within an image [Ear85]. One simple approach to automatic edge detection is to compare adjacent pixels against each other. Should a significant difference score exceed a certain threshold the pixel is considered an edge and is automatically marked as black otherwise it is marked as white. Many effective edge detection methods have been developed over the years including SOBEL [GW92] and the Canny edge detector [Can86]. Edge detection indirectly characterizes the shape of objects within videos and as a result is vital

for ongoing research into object detection and extraction which will inevitably improve the performance of video retrieval.

Texture is the term used to describe the different patches of an image that follow a particular pattern characterized by differences in the levels of brightness contained in an image. An image can have many different textures from an outdoor pool to a grass garden. Texture can be detected similar to the method used for the detection of edges, although a much lower difference threshold is used to detect texture during the comparison of adjacent pixels. Figure 2.6 gives an example of three different textured images. The detection of texture in images can aid in the identification of relevant documents during search but in addition texture detection can be used to aid in the development of specific higher level feature detectors, such as the beach detectors or water detectors.

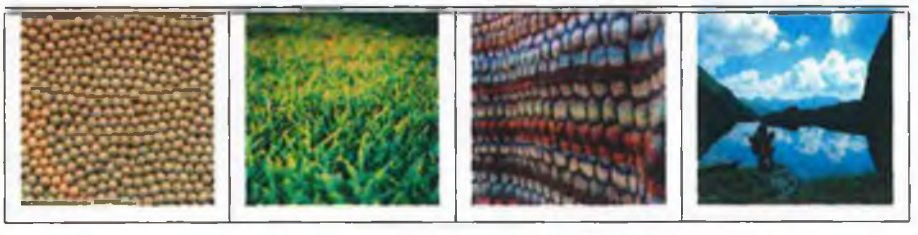

Figure 2.6: Example of 3 different textures Soybeans, grass, a jumper and an image containing many textures

The use of low-level features to aid in the retrieval of particular video shots has limited performance. Shots of "forest fires" or of "the Washington Monument" are accurately detected using the various low-level feature detectors, however consider the scenario where a user wishes to find shots of an "Airplane taking off". At present there is a large semantic gap between low-level features and their higher level meanings. The detection of concepts attempts to bridge this semantic gap and aid in the overall performance of a retrieval system.

Concept searching

Semantic information detection is very important within the retrieval community and particularly so given the visual nature of video retrieval, yet the

extraction of the semantic meanings from a visual image or video sequence is a highly complex task. In many commercial video libraries humans manually annotate content descriptions based on what they see within a shot, in the form of keywords such as "crime" or "violence", or with descriptions written by experts giving elaborate descriptions on the visual content or the emotions that are depicted within the video. Visual perception is subjective which can result in a lot of ambiguity. In addition to this, is the large consumption of valuable time and manpower needed to annotate a video. This makes manual annotation unattractive from the point of view of accurate, effective and efficient video manipulation and retrieval. Recently TRECVid in an attempt to bridge the gap between the low level features that are currently detected and the higher level semantic knowledge that exists in visual images, have built ontologies of high level features/concepts encouraging research into the development of automatic concept detectors which can be applied off-line to video collections subsequently enhancing an overall retrieval performance, should a user request contain a particular concept that is within the ontology.

Naphade [NS04] give an overview of a number of detection approaches that have been undertaken over the last few years. Each shot within a video collection is assigned a confidence score for each concept detector and is integrated into the MPEG7 [Com02] description of a video. However as mentioned earlier there are great challenges in creating concepts for broad domains such as TV news. It has become necessary to establish a standardised ontology a subject which is described in greater detail in Chapter 4.

Fusion

A user may query a video retrieval system using a combination of text descriptions and one or more visual images or video clips containing an example of their need. As we have seen in the previous sections, video has many features which can all be utilised to retrieve a particular shot satisfying a user's information need. These include automatic speech recognition, closed caption and optical

TRECVid 2002	1. Outdoor, 2. Indoor, 3. Face, 4. People, 5. Cityscape , 6.landscape, 7. Text Overlay, 8. Instrumental Sound, 9. Speech, 10. Monologue
TRECVid 2003	1. Outdoors 2. News subject face, 3. People, 4. Building 5. Road, 6. Vegetation, 7. Animal, 8. Female speech, 9. Car/truck/bus, 10. Aircraft, 11. News subject monologue, 12. Non-studio setting, 13. Sporting event, 14. Weather news, 15. Zoom in, 16. Physical violence, 17. Person x
TRECVid 2004	1. Boat/ship, 2. Madeleine Albright, 3. Bill Clinton, 4. Train, 5. Beach, 6. Basket scored, 7. Airplane takeoff, 8. People walking/running, 9. Physical violence, 10. Road
TRECVid 2005	1. People walking/running, 2. Explosion or fire, 3. Map, 4. US flag, 5. Building exterior, 6. Waterscape/waterfront, 7. Mountain, 8. Prisoner, 9. Sports, 10. Car

Table 2.1: Concepts assigned by TRECVid since 2002

character recognition texts within the text component. Within the visual component we can utilize low-level features such as colour, edge, and texture and higher level concept detections such as outdoors, cityscape, audio and road. All these parallel retrieval component results must be combined or fused in a way that should improve upon the performance of the best individual retrieval result, providing a user with the maximum overall retrieval performance and these are usually grouped into early fusion and late fusion methods. "Early fusion" combine multiple features into a single vector representation. "Late fusion" methods [MS05], fusion of individual feature scores occur once similarity matching has been performed for each of the individual features, are the most effective method of combining multiple feature retrieval streams and include method such as CombSum, CombMNZ, CombWtScore, Borda fuse and round-robin which are explained in detail in McDonald [MS05]. McDonald [MS05] examined various fusion methods that have been successfully implemented in traditional text on the various retrieval streams within a video retrieval model.

It was reported that CombSumScore, the summation of the normalised scores from the top N results (the traditional Combsum method), works very effectively for combining a single visual feature over multiple visual examples. It has been found that the weighted average of the normalized scores of the top N documents, achieve the best performance when combining text and visual results for a user's text and example image query.

2.3.1 The User Interface

The design of an interface is quite complex as it must contain sophisticated interface elements, to allow a user effectively search for a required video while at the same time display the results in a coherent manner, which allow the user to browse and navigate their way through the relevant result set. To achieve an efficient interaction within the system, the user interface must be easy to learn, simple and straightforward [GSG+03].

Figure 2.7: Interface for a video retrieval system - TRECVid 2003

Figure 2.7 shows an example of the main interface of a video retrieval system. The user may formulate queries using either a text string, an image or a combination of both text and image example. The user initiates the search by

clicking the search button which presents a ranked list of retrieved "Group of Shots" and their associated ASR transcript portions containing the highlighted query terms to the user on the right side of the screen. The "Group of Shots" are displayed such that the shot with the highest relevance is presented in the middle surrounded either side by two shots directly preceding it and two shots directly succeeding it in that particular video program. This allows the user to see the context within which this shot appeared in the video program. Each shot in the result set has associated with it:

- Keyframe: Represents the visual content of the shot. Clicking on the image initiates the playback of the shot from this keyframe.
- ASR transcript: The portion of ASR aligned with the shot is displayed.
- Save Checkbox: This allows the user to indicate that the current shot is relevant to the topic.
- Add to Query: This button allows a user to reformulate a query using a particular image and its associated ASR transcript portion. The image has usually been identified as relevant or very similar to their information need. By clicking this button the image and text are updated in the box directly below the text search box. To re-query the user must press the search button.

A searcher is also given the facility to browse an entire video programme by clicking on the "Browse this Programme" button (Figure 2.8). This displays the entire video represented by the keyframes of each shot. At the top of the content browser the user is supplied with a graphical timeline, which displays areas of the video programme, which match the user's query over a certain threshold. Clicking on any part of the timeline displays keyframes representing that part of the video content.

Figure 2.8: Browse this program facility of a video retrieval system - TRECVid 2003

2.4 TREC: A brief history

In the 1992 the first TREC (the **T**ext **RE**trieval **C**onference) conference was organised by the National Institute for Standards and Technology or NIST, a US government organisation, as part of an evaluation for DARPA's TIPSTER program [Har92]. Twenty five research groups participated in the first ever conference focusing on ad hoc retrieval and analysis of two gigabytes (GB) of text using fifty topics. This was a significant undertaking in 1992 as many systems were then unable to store 2GB of data. On an annual basis since TREC's establishment, research groups have an opportunity to evaluate their progress in the designing and implementing of information retrieval systems (of both text and of a multimedia content) using standardised guidelines and common evaluation procedures. Over the last few years as technology has improved and collections have increased in size, complexity and availability, a

number of retrieval challenges have been identified and investigated over several domains and specialised research areas such as text, video, spoken documents and cross lingual, through specific specialized tasks called tracks within the TREC forum. These tracks help stimulate interest in the research topic pushing state of the art in the IR field [Voo04].

2.4.1 TRECVid

TRECVid, the video track within the TREC conference, was introduced to focus attention on and evaluate research in content-based retrieval from digital video information [Hom05]. Since its establishment in 2001, TRECVid has evolved rapidly and is now a stand alone separate activity to TREC. There has been an increase from ten research groups participating in 2001 to forty one in 2005, an increase in the collection sizes from eleven hours of video in 2001 to two hundred and twenty hours in 2005 and an increase in the number of tasks from two in 2001 to five in 2005. These tasks include shot boundary detection and interactive and manual video search introduced in 2001, feature detection introduced in 2002, news story segmentation which was run in 2003 and 2004, fully automatic search facilitated in 2004 and 2005 and the BBC rushes which was introduced in 2005 [SO02, SKO03, SO03, SKO04a, SKO04b]. The shot boundary task encourages groups to refine the detection of boundaries such as "cuts", "fades" and "dissolves" with high accuracy within a video sequence. The feature detection task encourages groups to automatically identify specific concepts from within a video collection and these can be seen in Table 2.1. This task allows groups to research innovative ways of integrating concept detectors into video retrieval systems in an attempt to improve the overall retrieval performance. The news story segmentation task encouraged research into identifying different semantic news stories within a news broadcast. The interactive search task was introduced to evaluate the performance of video retrieval systems, by analysing a searcher's ability to effectively and efficiently search through a large video corpus in search of a particular topic. The manual

search task specifies that once a topic has been formulated by a professional, the system must perform the retrieval of relevant documents automatically. The fully automatic search is similar to the manual search task, however the official topic description are unmodified and the system preforms the retrieval of relevant documents automatically. Finally the BBC rushes task was introduced in 2005 to investigate ways of searching through material which unlike previous video data used so far, is unstructured, unprocessed, and contains little or no metadata.

2.4.2 Collection

In order to improve and encourage research in the area of information retrieval, it is necessary to build large data collections that model as close as possible a real world data collection, upon which research work can be carried out. The subsequent results and findings of experiments on the model collection could then in theory, perform similarly on a real world scenario. As in the main TREC conference, participants of TRECVid are supplied with a set of documents which in this case consists a collection of digitalised video and a set of predefined topics. In addition participants of TRECVid are also provided with supplementary data including Automatic Speech Recognition (ASR) transcripts supplied by LIMSI [JGA02] and shot boundary evidences detected by the CLIPS-IMAG group. This common shot boundary detection data facilitates a common unit of retrieval in referring to a particular video segment and thus allows easier cross comparisons to take place between different systems. The video collection used in TRECVid 2001 consisted of 11 hours of selected video taken from the NIST Digital Video Collection Vol-1[Nis] and the open-video project collection [Mar01], which consisted of selected NIST projects and U.S. government documentaries respectively, dating from the 1980's to early 1990's. The video collection in TRECVid 2002 was larger compared to 2001, consisting of seventy two hours of digitalised video. The collection was segmented into two sets. Forty two hours were assigned as test data with the remaining thirty

hours designated for training and development. The collection comprised of digitalised video acquired from the Prelinger Archive [Arc02], the Open Video Project and some stock shot videos which were provided by the BBC Archive. The videos provided by the Prelinger archive dated from the 1930's to early 1990's and varied in content from educational to advertising to industrial and amateur footage. The visual quality of this collection was poor containing various encoding abnormalities which resulted in systems performing quite poorly during their various experiments. TRECVid 2003 attempted to model a real world news collection. Although the collection was smaller by comparison to an operational news collection such as the Físchlár news collection which contains over two years or 250 hours of news footage [SGL+04], the goal was achieved and the search test collection consisted of one hundred and twenty hours of televised news programmes. The collection consisted of news programmes following similar evening news formats broadcast during 1998 from two U.S. channels, ABC and CNN and thirteen hours of CSPAN news, which broadcast debates of the US Congress in 2001. Both the visual and audio quality of the collection was significantly better than previous years. The TRECVid 2004 search test collection was very similar to the TRECVid2003 collection consisting of seventy hours of news broadcasts from ABC, CNN and CSPAN. The TRECVid 2005 collection consisted of 169 hours of digitised video divided into 74 hours of ABC, CNN and 43 hours of Arabic and 52 hours of Chinese televised news from 2004 collected by the Linguistic Data Consortium.

2.4.3 Topics

TRECVid topics are generated by NIST to model as close as possible a real-world request to a video collection. They are based on real queries found in professional visual retrieval environments which are classified into various types such as person thing, event and place [AE96]. A typical TRECVid topic is composed of a short textual description and an optional example video clip, image and/or audio example of the topic. Figures 2.9 shows a topic used for

TRECVid 2004 and which was composed of a short text description, image and video example. The relevant shots for those particular queries are displayed directly under the example images. It should be noted that a keyframe for a specific shot may not be relevant to the topic even though the shot the keyframe represents is relevant. Potential TRECVid topics are created manually by watching video from the respective test collection without audio. The absence of audio ensure the non-biased generation of the text description for the topic [SKO03]. In TRECVid 2002 this restriction was not a requirement and consequently words from the audio track inevitably appeared in the short text description of the topic leading to enhanced performance when searching using ASR-based retrieval [SO02]. The visual examples to accompany the topic descriptions were chosen separately to the relevant shots for that topic in the collection, again to avoid potential biases for results in the visual domain. As well as modeling the topics on real user requests, composers were also expected to solely generate topics that were not too difficult, had multiple relevant shots and if at all possible from across multiple videos. A full list of the topics used in TRECVid 2003, TRECVid 2004 and TRECVid 2005 are available in Appendix A.

Figure 2.9: Example images for Topic 125: "Find shots of a street scene with multiple pedestrians in motion and multiple vehicles in motion somewhere in the shot."

2.4.4 Relevance Judgements

The relevance judgements or "qrels" are another main element of the TRECVid structure. They are the ground truth[1] of "right answers" or relevant shots to a particular topic. Each participating research group run the TRECVid topics against the video collection using their individual video retrieval systems and send their ranked lists of results back to NIST for evaluation. The submitted results from all groups are pooled together for a given topic, duplicates are removed and the merged list is manually assessed down to some fixed depth, in rank order, for relevant shots thus creating a ground truth of all known relevant shots. This enables comparative assessments among all the submitted results. An entire shot is viewed to determine its potential relevancy as opposed to viewing only the representative keyframe. Relevancy of a shot is a binary decision (relevant or not relevant), if an image anywhere in the shot contains information however small on the topic then the shot is considered relevant. These relevance judgements however are incomplete as pooling is used.

2.4.5 Evaluation Measures

The evaluation measures determine the effectiveness of the video retrieval system to accurately retrieve relevant shots for a specific topic. A systems performance measure can be calculated over an average of all topics within a submitted run or per individual topic.

There are many evaluation measures each looking at the results from a different perspective and chosen depending on a user's preferences for how he/she would like to assess the performance of the system. The traditional standard measures of recall and precision form the basis of the evaluation measures used within video retrieval. Higher precision as opposed to recall is in the majority of cases the preferred outcome of a retrieval system as it tells a user how accurate the

[1] this is the definitive set of relevant documents

results are to the topic. Higher recall and hence less precision is sometimes beneficial if a user is researching information on a particular topic and is looking for as many documents as possible on the general area. Three other evaluation measures to evaluate effectiveness of a retrieval system include Average Precision, Mean Average Precision, Precision at cut-off points and Interpolated Precision. More detailed information on precision and recall, average precision and mean average precision can be seen in Chapter 1 section 1.4.

Most video retrieval systems display between 10 and 30 of the top ranked results within an initial browser window. The ***Precision at cut off*** points measure, assesses the precision of a list of results at various cut off points for example after 10, 20, 30, 40 documents in a list. This allows the evaluation of the quality of the search results that are displayed on the first page of a users interface.

The number of documents marked as relevant to a topic will vary from topic to topic. In order to allow comparative analysis over a set of topics, a measure called ***interpolated precision*** is used. This creates a set of eleven standard recall points starting at 0.0 and increasing by increments of 0.1 to 1.0. Precision values are then interpolated to this standard range. So at any one of the standard recall values for each topic and for any system we have a precision value. This precision value is calculated using a rule which specifies that the precision values for a standard recall value n (where n is a number) is equal to the maximum precision value actually achieved for each recall value greater than or equal to the recall value n. While there is no precision value for recall value of 0.0 interpolated precision assumes the rule and assigns the maximum precision value actually achieved for the nearest actual recall value. Consider for example actual recall values of 0.35, 0.8 and 1.0 illustrated in Figure 2.10. Any standard recall value from 0.0 up to .35 (0.3) is assigned an interpolated precision value corresponding to maximum precision value achieved at that actual recall point. Similarly all the standard recall values ranging from 0.4 to 0.8 are assigned the interpolated precision values corresponding to the maximum precision value achieved at the actual recall value 0.8 etc.

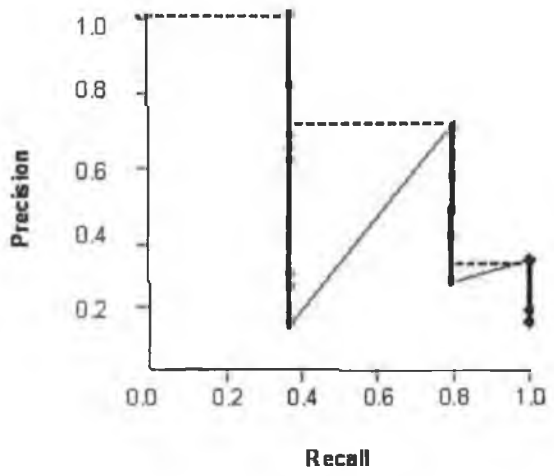

Figure 2.10: Interpolated Precision Graph

2.5 The current state of video systems

Video retrieval is a relatively new area in Information Retrieval research. As video collections are far more complex than traditional text collections in terms of the variety of components making up a typical video sequence, sophisticated multimedia systems such as a video retrieval system require searching techniques and interface elements. In this section we will describe three state of the art video retrieval systems namely the Físchlár Digital Video Library, the Informedia Digital Video Library and the Marvel Multimedia Analysis and Retrieval System.

2.5.1 Case Study- Físchlár Digital Video Library

In order to showcase and evaluate on-going research into digital video indexing, browsing and retrieval the Center for Digital Video Processing (CDVP) in Dublin City University developed a suite of web-based interactive video search/browse system called Físchlár. Físchlár TV, the first version of the Físchlár family, was implemented and shared within the entire university campus [OMM+01]. This system was designed to mimic the actions of a VCR set

Figure 2.11: Físchlár TV browse and playback displaying hierarchical browser interface

up, whereby a user could select a specific broadcasted program from an on-line TV schedule for recording. At the specified program broadcast time the system automatically captured and encoded the program into MPEG1 format. Shot boundary detection and keyframe extraction modules were then executed upon the digitalised video. The resulting set of representative images were displayed through a number of various browser interfaces allowing the user to search through and playback his/her selected content from any point in the video [LSO+00, LSM+01]. The interface is divided into two main functions. On the left hand side of the screen the user is presented with a list of videos available for browse and playback. Clicking on any of these videos will display an overview of that video on the right hand side of the screen. Figure 2.11 illustrates the hierarchical browser interface.

Físchlár News [SGL+04], a variation of Físchlár TV was designed to support

research on the news broadcast domain. Figure 2.12 shows an example of Físchlár News. This system recorded daily broadcast news and enabled a user to search through the recordings for specific topics of interest using keywords that are matched against the closed caption text extracted from the news programs. On the lefthand side of the Físchlár News screenshot we see a search box and a calender. Clicking on any date displays a list of news stories that were presented within that news broadcast while the righthand side of the interface shows the video. Browse and playback options were available by clicking on any story of interest.

Figure 2.12: Físchlár News interface

On an annual basis as a direct result of participating in the TRECVid conference, the CDVP developed a variation of the original Físchlár video retrieval system purposefully designed for conducting the search task of each particular year. Figure 2.13 displays an interface of Físchlár TRECVID2003. Físchlár TRECVID2003 was designed to accommodate object detection and relevance feedback in an attempt to improve the overall search and retrieval of video

Figure 2.13: System developed for TRECVid 2003

content consisting of news broadcasts from ABC, CNN and CSPAN.

2.5.2 Case Study- Informedia Digital Video Library

The first Informedia Digital library, Informedia-I, was developed in 1994 by Carnegie Mellon University (CMU) integrating various aspect of research being carried out within their group on multimedia understanding. They used shot boundary detection and keyframe detection to segment the videos and a software package Sphinx which automatically transcribed the audio track from the collections consisting of radio and TV news and documentary broadcasts. They differ to the Físchlár system by performing search and retrieve against the ASR, text overlay and closed caption texts as opposed to just the ASR and closed caption texts. They introduced the concept of *video skimming* which allows a user to view a particular video of interest rapidly without having the added noise effect of the common fast forward of frames and audio track.

Informedia II (Figure 2.14) focuses not only on the retrieval of video content but also on the greater understanding and access of video content though summari-

sation and visualisation. They integrated techniques that identify important persons, places, dates and time references from the rich metadata associated with the video content. They incorporate geographical thesauri to cluster related stories and documentaries according to their geographical region. Navigation of these stories clusters is achieved via an interface that incorporates a map of the world and timeline bar. More information on the Informedia project can be accessed at the Informedia home page http://www.informedia.cs.cmu.edu/

Figure 2.14: The Informedia II multimedia search engine[a]

[a]Picture taken from the informedia Home Page http://www.informedia.cs.cmu.edu

2.5.3 Case Study- Marvel Multimedia Analysis and Retrieval System

IBM research developed a video retrieval system, MARVEL Figure 2.15, which showcases the group's research into semantic and feature based searching as well as the conventional and traditional text based searching through a video collection. The system is composed of two components.

The first component, the MARVEL multimedia analysis engine, automatically labels or annotates video content thereby reducing costly manual annotation. Automatic annotation is achieved though machine learning techniques that consider visual, audio and text components. This goes some way toward bridging the gap between low level feature detections and their higher level semantic meanings.

The second component, the MARVEL multimedia search engine, enables the user to search the collection of video using a number of options such as semantic-based queries (text queries or specific keywords that are part of an ontology). These are matched against the rich semantic data of MPEG7 annotations generated by the multimedia analysis tool, feature based queries enabling a user to issue example image queries or video clips which are matched against images within the collection though the use of MPEG-7 feature descriptions and finally conventional text based queries which are matched against the ASR transcripts, closed caption, text overlay and the MPEG7 annotations.

In this section we described three typical video retrieval systems within the video information community. They would typically incorporate novelty detection modules as part of a their search architecture in the future.

2.6 Video Annotation

As we have seen within the last few sections of this chapter there is a large difference between the text and visual medium. Highly semantic information is integrated naturally within text documents and is easily extracted by computers for the accurate retrieval of relevant data using various text matching techniques. However within a video document the semantic information is implicit for the visual components and occasionally explicit in the spoken audio text. At present the automatic extraction of semantic information from visual media is a very difficult task, yet it is a resource which can aid in the retrieval performance of a video retrieval system. Once viewed, humans have the ability

Figure 2.15: Screengrab of the MARVEL multimedia search engine[a]

[a]Screengrab taken from the IBM Research home page http://www.research.ibm.com/marvel/

to accurately perceive and understand the semantic information of visual media. As a result the manual annotation of video content is a logical solution to the description of semantic information displayed within the content of an image. However the annotation of visual content is very subjective and this can cause a lot of ambiguity as humans perceive visual content differently depending on when or where it was completed and also on the different factors that make up an annotator's personal background. For example where a person was born, family, religious customs, friends and education all have an influence on how a person will perceive an event or thing. This leads to the same video sequence being annotated with different descriptions and subsequently indexed differently. Manual annotation is also very costly in both time and manpower.

Annotation
outdoors, river, boat, buildings, people

Figure 2.16: Annotation of an image

If we consider the image in Figure 2.16, it is obvious that very different concepts can be used to describe its contents. The example annotation for the image does not state that the image is set in Venice or that the image contains two moored gondolas. Some methods that can be employed to reduce the amount of ambiguity when annotating video content include; the definition of standard guidelines, the training of annotators and the definition a standard list of concepts from which the annotators must describe an image. Video annotation is an important part of this thesis and we shall return to it later.

2.7 Summary

In this chapter we have introduced multimedia information retrieval. We have seen that the retrieval of video data is much more complex than that of traditional text data. Within digital video we have detailed the many media components that need to be considered during the manipulation of video content. We have outlined that major challenges that exists within the video information retrieval community including the enormous size of video data, the visual representation of retrieved search results to the users, the lack of accurate feature detectors and the inability to automatically understand the semantic meaning within the visual media component of video. We investigated the individual components that are necessary within a video information system.

Chapter 3

Introduction to Novelty

In this Chapter we will introduce the concept of novelty detection in information retrieval. We will look at the different types of novelty detection and then outline the assumptions that are necessary in identifying a "novel" document. We will then introduce the TREC novelty track and lastly describe our novelty detection model developed for the novelty track in 2004.

3.1 Novelty Detection

To clearly illustrate the idea and directly motivate novelty detection in information retrieval we consider the following analogy. A child is given an essay as a homework assignment and requires as much information as he can possibly get on the topic but is restricted by the amount of time he can spend searching through the data collection. He enters a keyword into the retrieval system best describing his information need and hits the "search" button. The list of documents returned are ranked in descending order of their "degree of relevance" to the request. The child reads the first document and gains knowledge on the topic. He returns to the ranked list and clicks on the next document that was returned as relevant to his topic. He reads the document looking for

new information to add to the knowledge he has already gained from the previous document, however he does not acquire any new information from this document. This is an example of a redundant document. Reading the same information has no incremental value when trying to increase ones knowledge about a specific topic.

During the last two decades we have seen significant improvements in technology with capabilities to create, capture and store vast amounts of information effectively and efficiently. It has become imperative that methods are developed that allow users to quickly and effectively sift though this vast sea of information and focus on the particular information they require. Within information retrieval a user submits a query and receives a list of documents that are potentially relevant to the request. If the list of documents retrieved is quite small then ordering by degree of relevance seems logical, as the user can quickly determine what documents will suit his/her information need. However in the majority of cases when a user issues a query, the user is presented with a large list of documents each with a high potential of containing information that will be useful to the user. It is highly possible, particularly within the news domain, that a user will see information contained within a document that they have already seen in a previously read document, which may have been phrased differently or presented in a different manner. In the event that such a scenario occurs a user has gained no new knowledge and has wasted time and effort. Novelty detection aims to reduce the amount of redundancy within a results set, by identifying new information to present to the user. It challenges the traditional methods of ranking documents by maximal degree of relevance to a query [Sal89] by identifying whether or not these documents contain new information to a particular users query. Novelty detection is defined as the detection of documents that provide "new" or previously unseen information. "New information" in search result list is defined as the incremental information found in a document based on what the user has already learned from reviewing previous documents in the document ranking. It is assumed that as

a user views a list of documents, their information need changes or evolves, and their state of knowledge increases as they gain new information from the documents they see. The automatic detection of "novelty", or newness, as part of an information retrieval system could greatly improve a searcher's experience by presenting "documents" in order of how much extra information they add to what is already known, instead of how similar they are to a user's query. This could be particularly useful in applications such as the search of broadcast news and automatic summary generation. Broadcast news is abundant with information repetition as stories reappear over time. The use of novelty detection could identify new unseen information about a story and display a list of novel documents to the user. The occurrence of redundancy within a summary defeats the purpose of a summary, consequently interest in novelty detection has increased in the research area of automatic summary generation with many systems now containing novelty detection modules in an attempt to generate non redundant summaries of a document or of multiple documents. This interest has mainly concentrated on finding better ways to detect novel or new sentences, as they are usually more informative and hence of most importance for inclusion in a summary [CG98].

3.1.1 Definitions

Novel information is new information not previously seen in any other document so far.

Redundant information is information within a document that has been seen within relevant documents that have already been presented to a user. The term "redundant information" constantly appears during the course of novelty investigation. From this point forward we refer to non-novel documents as redundant documents. Likewise we refer to non-redundant documents as novel documents.

3.1.2 Assumptions

In order to avoid ambiguity, the identification of novel information is carried out under the following assumptions.

- **Assumption 1:** High precision is not always guaranteed when returning relevant documents to a users query for a particular topic. As a result and for simplicity we make the assumption that novelty detection is performed on a list of documents that are all known relevant to the users request.
- **Assumption 2:** The detection of relevant documents for a user's query within a data collection is a separate task to the detection of novel documents for a user's query within a retrieved set.
- **Assumption 3:** The novelty of a document is dependent on the documents that have been previously displayed to a user.
- **Assumption 4:** We assume a user is only tolerant of receiving information that he may already know due to some background knowledge he may have on the topic.
- **Assumption 5:** We assume that a user knows nothing about the topic at the time the initial document is displayed and that all knowledge about the topic is gained as a user progresses through a list. This means that the first document of any list of relevant documents will be considered novel. This is not quite reflective of the real world but it is an assumption that allows us to address novelty issues directly.

3.2 The History of Novelty in Information Retrieval

There are three main forms of novelty detection, each closely related while at the same time attempting to accomplish different goals. This similarity has

resulted in the migration of techniques and approaches across the three different novelty detection areas. The first form of novelty detection identifies new *"events"* across an entire collection of data. Events are defined as "something that happens in some specific time and place"[SC01], for example an explosion. New event detection is designed to automatically detect specific characteristics that could signal the presence of a new event. This kind of novelty detection aids a user monitoring a continuous news stream by indicating when something new is first reported such as a helicopter appearing in the horizon. This was called first story detection (FSD) and was initially investigated in a report written by James Allen et al [AJR+99].

The second kind of novelty detection focuses on returning new stories about known topics over an entire collection and is currently being researched and investigated within both the TREC filtering track and the TDT topic tracking and story linking detection tasks. Topic Detection and Tracking (TDT) [DT] is an annual benchmarking event that focuses research on event based organisation of broadcast news.

In this thesis we concentrate on the third type of novelty detection, intra novelty detection, which identifies novel information within a list of potentially relevant documents retrieved for any user-specific topic and the subsequent re-ranking of documents based on their degree of "newness". Intra novelty detection is carried out on a subset of the collection, the set of highly ranked documents, as opposed to the entire collection in event and topic tracking detection. It concentrates on the semantics found within the vocabulary and determines the amount of new information that is present within a document. The detection of new information is a relatively new research area. Prior to a paper by Zhang et al. [ZCM02] little research had taken place on the construction of mathematical models to represent intra topic novelty detection. This was partly due to a lack of evaluation data and partly due to the ambiguity of the terms "novelty" and "redundancy". In their paper Zhang et.al focused on topic novelty detection in adaptive filtering, examining models previously applied to other areas such as

natural language processing and traditional information retrieval and adapting them to detect novel information. The approaches taken involved set difference, geometric distance and Cosine distance metric and a metric based on a mixture of language models, all of which utilised word frequency patterns to determine the novelty of the documents.

3.2.1 Summarisation

Multi-document summarisation is strongly related to the ideas in this thesis. The main purpose of a summarisation system is to highlight new and important information and decrease the amount of redundant information that is passed to the user. As a result many multimedia summarisation systems contain novelty detection modules. The best known work associated with novelty detection and the re-ranking of retrieved results is "Maximal Marginal Relevance" (MMR), presented by J. Carbonell et al. [CG98] in which the Cosine similarity of vectors is used to detect redundant information contained within a document. They introduced the concept of "marginal relevance". A document has high marginal relevance if it is both relevant to a user request and contains very little similar information when compared to the previously seen documents.

Allan et al.[AGK01] have investigated novelty detection on a TDT corpus through the use of different language models. Their work involves developing a language model to estimate the probability that a sentence is novel when compared to its predecessors using both individual and cluster sentence models.

The most recent activity within the novelty detection research, that is closely related to the work carried out in this thesis, has taken place within the TREC novelty detection track discussed in Section 3.4.

3.3 Approaches to Novelty Detection

Allan et al. [AWB03] have given a concise summary of various models used for novelty detection and experimented with them on the data used in the TREC Novelty 2002 (described in Section 3.4.2). These range in complexity from simple word counts, set differences and Cosine distance measures to language models using KL divergence with different smoothing techniques. The following is a description of some of the possible approaches that can be taken to determine the novel value of a document.

Document-Document Distance: This model illustrates how users or assessors prefer to investigate the novelty of a sentence, by comparing its similarity to other documents one document at a time rather than against the entire set of previously seen documents. The model measures redundancy, $R(d_t|d_i)$, based on the distance of the current document to a previously seen document. R will be high if $d_t = d_i$, that is if they are duplicates.

New word count: This simple approach assigns a value to a sentence based on the number of unique words it has, when compared to all other documents that have been seen in the collection and is defined by equation 3.1.

$$S(d_i|d_1..d_{i-1}) = \|A_{d_i} \cap A_{d_j}\| \tag{3.1}$$

$S(d_i|d_1..d_{i-1})$ represents the novelty score while A_{d_i} represents the set of words occurring in the document d_i. This approach was one of the best performing approaches in TREC Novelty 2002 [AWB03].

The Set difference: This is another set oriented approach, measuring the redundancy of a document by taking into account the frequencies with which a word can occur in a document. It tries to model the fact that a document with words occurring more frequently in it will most likely contain more information on that topic. However it also considers that a word may occur too frequently in a particular topic lending no useful information or being in a sense, a stopword.

The redundancy distance or similarity is measured on a document-by-document basis. The model described mathematically is given in equation 3.2

$$S(s_i|s_j) = |A_{s_i} \cap \overline{A_{s_j}}| \tag{3.2}$$

$$w_k \in A_{d_i} iff(w_k, d_i) > k \tag{3.3}$$

$$count(w_k, d_i) = (\alpha_1 \times tfw_k d_i) + (\alpha_2 \times dfw_k) + (\alpha_3 \times rdfw_k) \tag{3.4}$$

where

S is the measure of redundancy or similarity score, s_i, s_j are the documents A_{s_i} , A_{s_j} are the set words in s_i and s_j respectively.

and

$tfw_k d_i$ is the number of times the word w_k appears in the sentence d_i.
dfw_k is the number of documents not relevant containing the word w_k.
$rdfw_k$ is the number of sentences previously seen containing the word w_k.
α_1, α_2, α_3 and k are all parameters chosen according to the information being trained.

The Cosine Distance: This approach models a document as a vector in an m-dimensional space, with each unique word representing one dimension. The weights assigned to each word are determined using the $tf * idf$, weighting algorithm. The redundancy measure is calculated on the negative of the Cosine angle between two document vectors. It is defined in equation 3.5.

$$S(d_i|d_j) = Cos(d_i, d_j) \tag{3.5}$$

Or

$$NScore = -\left(\frac{\sum_{k=1}^{n} w_k d_i \times w_k d_j}{\|d_i\| \|d_j\|}\right) \tag{3.6}$$

This approach works well when defining novel scores for full documents however the performance decreases on documents of a shorter length [AWB03]. The reader is directed to Allan et al. [AWB03] for a more detailed discussion on

each approach including experimentation and results.

There have been several complex approaches that attempt to measure the novelty of a document by measuring the difference in word distributions. Although more complex, these approaches have not produced any significant improvements over other approaches previously described within the TREC Novelty Track evaluations. These approaches include language modeling approaches which are currently very popular in experimental IR. The interested reader is directed to the PhD dissertation by Djoerd Hiemstra [Hie01] for more information on language models [SH03].

3.4 The TREC-Novelty Track 2002-2004

For three years in a row (2002-2004), the annual Text REtrieval Conference (TREC) ran a novelty track [Har02, SH03, SH04]. The overall goal of the novelty track was to challenge the traditional method of ranking by degree of relevance to a user's query by exploring, encouraging and evaluating methods that identify new information within documents and subsequently reduce the amount of redundant and duplicate information, which is displayed to a user for a specific topic.

Within the novelty track participants are given an opportunity to create systems that automatically retrieve relevant documents for a specific topic in addition to creating systems to automatically retrieve novel documents from a predefined list of relevant documents. The identification of relevant documents is a separate task to the detection of novel documents. In this thesis we are interested in the novelty detection task. Participants of the novelty task in the novelty track were given a list of search topics and an ordered list of relevant documents associated with each specific topic. Participants were required to find the documents that provided "novel" information to the user. The track is based on the detection of novel information, at "sentence level" as opposed to full document text level.

3.4.1 Evaluation Measure within TREC-Novelty

Within the Novelty track the evaluations of the detection of relevant and novel sentences were assessed separately. The track worked with an unranked list of documents and as such traditional evaluation measures such as precision at cut off points for example precision at 10 or mean average precision cannot be used. It has been observed over the various filtering tasks within the TREC community over the last few years that set recall and precision do not average well and as a result can lead to a misleading representation of a systems performance [SH03]. As a result the "Fmeasure" (equation 3.7) defined as the harmonic mean between recall and precision is the primary measure of effectiveness within the novelty track. This measure evaluates the quality of the documents returned within the set. It is based on van Rijsbergen's, E-measure, a function of set recall and precision [vR79]. It contains a parameter β, which determines the relative importance of both precision and recall. A β value of 1 indicates an equal emphasis on recall and precision is used within the novelty track [Har02, SH03, SH04].

$$F_{\beta=1} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3.7}$$

Or

$$\frac{2 \times |Rel(NEW) \bigcap Ret|}{|Ret| + |Rel(NEW)|} \tag{3.8}$$

This measure however is not accurate in cross system comparisons as an Fscore can be achieved using many variations of recall and precision values [SH03, SH04]. As a result a system achieving an Fscore of 0.6 may not perform the same as another system also achieving an Fscore of 0.6 as an Fscore of 0.6, clearly seen from Figure 3.1, can reflect a range of precision and recall values for each system. Variations in the Fscore can be due to a variation in either precision or recall or both.

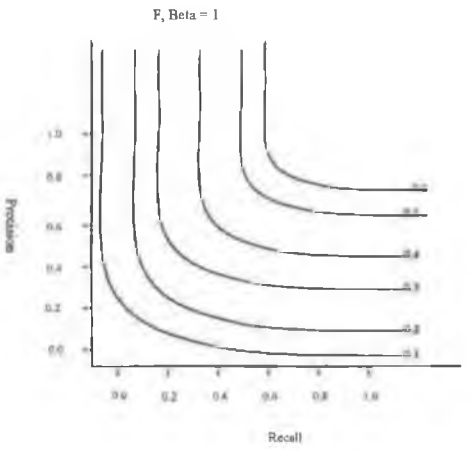

Figure 3.1: The F-measure Graph plotted in precision- recall space. The lines show the contours at intervals of 0.1.

3.4.2 TREC Novelty 2002

The first novelty track was initiated in 2002. The following section describes the collection and task used and then we describe the successful approaches to novelty detection in 2002.

Collection

The collection of data used for the Novelty Track in 2002 consisted of government documents selected from previous TREC collections namely TREC's 6, 7 and 8 [Har02]. The track selected fifty topics from the set of a hundred and fifty topics used in the previous tracks altering the original TREC topic statement to include a description tag which indicated the accessor's information need during the manual construction of the ground truth data for the respective relevant and novel sets. All documents were assessed for their potential relevance with twenty five documents manually selected and ranked according to their degree of relevance to each of the fifty topics. Each document was automatically segmented into its individual sentences at which point each sentence was assigned

a unique identifier. Each sentence was then manually assessed for relevance to the topic and ranked accordingly. The assessors were then required to assess the relevant documents in their ranked order and select a subset of the sentences that they deem novel with respect to previously seen sentences.

Tasks

Participants were given the set of topics and their corresponding relevant sentence-segmented documents in rank order and asked to firstly automatically determine the relevant sentences for each of the fifty topics. Secondly using their ranked list of relevant sentences, participants were required to automatically determine a subset of sentences from within their relevant set of sentences that provided new or novel information with respect to previously seen sentences in the list for the particular topic. The ranked order of relevant sentences was observed at all times. Results were submitted to NIST for manual evaluation.

Successful approaches to novelty detection 2002

Table 4.3.2 shows the Fscores of the two best performing approaches to novelty detection in 2002 undertaken by Tsinghua University and Queens University against the baseline of randomly choosen novel sentences. Tsinghua University employed sentence expansion and used an overlapping measure to determine the novelty score of a sentence depicted by equation 3.9.

$$novelty_{ts} = \frac{X \cap Y}{Y} \tag{3.9}$$

In this equation, X is the sentences previously seen and Y is the current sentence being investigated.

Queens University approached the task using traditional information retrieval methods treating documents as sentences. They used a novelty coefficient based on Jaccard's coefficient which takes two sets X and Y containing the terms occurring in the two sentences and determine the novelty of a sentence based

Group	Fscore
Random sentences	0.036
Tsinghua	0.217
Queens	0.193

Table 3.1: Best performing group Fscores against random chosen novel sentences

on a predefined threshold value (3.10).

$$novelty_q = \frac{|X \cap Y|}{|X \cup Y|} \tag{3.10}$$

Summary of the novelty track in 2002

It was observed that there was very little redundancy among many of the relevant sentences due to the nature of the data collection which included government documents from a sequential time period. This was rectified in the Novelty Tracks of 2003 and 2004 by using a data collection consisting of several news sources captured during an overlapping time period. This increased the redundancy of information within the collection. It was also observed that the detection of novelty is somewhat harder than the detection of relevancy.

3.4.3 TREC Novelty 2003-2004

The following section describes the collection and the topics given to all participants. We will then describe the successful approaches to novelty detection in the subsequent years 2003 and 2004. The Novelty Track concluded in 2004.

Collection

Participants of the Novelty track from both 2003 and 2004 were provided with a collection of documents from the AQUAINT collection. This collection con-

tains news documents from three different newswires sources, The New York Times Service, Associated Press and the Xinhua News service, all taken from an overlapping time period (1996-2000) [SH03][SH04]. The reason for using three sources of material was to increase the likelihood of near-duplicate or redundant news articles occurring across the different newswires thereby increasing the realism of the experiment. Fifty topics were constructed with a total of twenty-five relevant documents per topic collected prior to the release of the corpus. These topics were divided equally into two types: "event" topics which focused on a particular event that occurred within the time period such as the launch of a space craft, and "opinions" where topics focused on the different points of view on particular issues such as the war in Iraq. Of the fifty topics, twenty eight topics had relevant documents from the three sources NYT, AP and Xinhua and twenty one had relevant documents from two sources.

Documents were ordered chronologically rather than according to their degree of relevance to the topic which had occurred in that novelty task of 2002. As mentioned earlier, TREC evaluated novelty on a sentence level and as such, the relevant documents were broken into sentences. Each sentence of approximately twelve words was given a unique identifier, assessed for relevancy to the topic and consequently placed in the appropriate relevant or non-relevant sets. The assessors were then required to examine the relevant sentences in order and select a subset of these sentences containing novel or new information on the topic. In 2004 there was a slight change to the corpus provided to the participants. For the fifty topics constructed in 2004, each topic had twenty five relevant documents similar to 2003 but also had an additional zero or more non relevant documents assigned to them.

The chronological ranking was introduced in an effort to overcome the problem of which document should be displayed first. The theory was that in news documents, background information is usually given more completely in earlier reports and is repeated more briefly later on as new news is added to the report.

Tasks

Four tasks were defined for groups participating in both the novelty tracks of 2003 and 2004 which included:

1. Researchers were given the set of relevant documents for each topic and were asked to automatically identify all the relevant sentences for that topic. They were then required to automatically select a subset of these relevant sentences that provided novel information.

2. Researchers were provided with the relevant sentences in all the documents and asked to automatically identify the novel sentences.

3. Researchers were provided with the relevant and novel sentences in the first five documents only and asked to automatically provide the relevant and novel sentences from the remaining relevant documents. This task was slightly different in 2004 where some topics may not contain any relevant or novel documents due to the addition of non-relevant documents into the collection.

4. Researchers were provided with all the relevant sentences from the topics and novel sentences from five of the documents. Their task was to find the novel sentences in the remaining set of relevant sentences.

Successful approaches to novelty detection 2003

In this section we present the two best performing approaches to novelty detection in 2003, undertaken by the Chinese Academy of Science and the National Taiwan University. The Chinese Academy of Science achieved an Fscore of 0.819, approached the novelty task by defining a new algorithm called the "new information degree" which measures the novelty of a sentence compared to previously seen sentences. Analysis of the novelty of a sentence is carried out on a sentence by sentence basis rather than assessing the novelty of a sentence

against a set of previously seen sentences. There were two variations of the new information degree. The first analysed the *idf* values from both the collection of sentences seen to the current point in time and the current sentence under investigation. It was defined as

$$NID_1 = 1 - \frac{\sum idf_{bs}}{\sum idf_{cs}} \tag{3.11}$$

where idf_{bs} are the *idf* values of words appearing in both sentences and idf_{cs} are the *idf* values of words appearing in the current sentence.
The second analysed the bi-gram word sequences (i.e taking two words at a time from a sequence of words) between the current sentence s_n and the previously seen sentence s_{n-1} defined as

$$NID_2 = 1 - \frac{|bi_{matched}|}{|bi_{all}|} \tag{3.12}$$

where $bi_{matched}$ is the number of bi-gram words matched between the sentences s_n and s_{n-1} and bi_{all} is the total number of bi-grams word sequences occurring in the current sentence under investigation. They used a static threshold which determines whether a sentence is novel [JZX03].

The best run submitted by the National Taiwan University achieved an Fscore of 0.812. Their algorithm attempted to differentiate the meaning of a sentence by utilizing the reference corpus to expand the sentences. Sentences whose similarity with the set of previously seen sentences exceeds a predefined static threshold are considered redundant, otherwise the sentence is novel and put into the set of seen sentences [THC03].

Successful approaches to novelty detection 2004

In 2004 the best performing systems included our own CDVP/DCU submission, Meiji University and University of Massachusetts. Meiji University considered the rarity of words in a sentence to determine the sentence's novelty value,

achieving an Fscore of 0.619. They used a combination of three methods, "the Redundancy score", "the Sentence Weight Score" and "the Scarcity score". The "Redundancy score" estimates the redundancy of a sentence by finding its similarity to all the sentences which have already been identified as novel using the cosine similarity measure. The "Sentence Weight Score" measures the novelty of a sentence by assessing the rarity of a word within a *small range* of sentences previously seen, defined using *Nwindow_idf* which is the document frequency in the past N documents (see equation 3.13). The Sentence Weight Score is defined as

$$SentenceScore(s) = \sum_{i} tf(w_i) \times Nwindow\_idf(t_i) \tag{3.13}$$

where $tf(t_i)$ is the frequency of the word t_i in the sentence s, *Nwindow* is the number of sentences previously seen. $Nwindow\_idf(t_i)$ is the inverse document frequency of the word t in the previously seen documents. The third measure the "Scarcity score", identifies unique or infrequently occurring words within a collection of sentences [KKK+04].

The University of Massachusetts investigated the novelty of a sentence to the previous sentences using the Cosine similarity measure between a sentence and its previous sentence. A sentence with a similarity measure above a certain predefined static threshold was considered redundant. They also considered the occurrences of new named entities including persons, location and organisation etc. A sentence with previously unseen named entities was considered novel. They achieved an Fscore of 0.618 [AJAC+04].

The CDVP/DCU submission described in section 3.5 achieved an Fscore of 0.622.

Summary of the novelty tracks in 2003 and 2004

A number of interesting observations were noted after the completion of the novelty track of both 2003 and 2004. It was observed during the comparison of the respective Fscores for novelty and relevance detection over the topics that the detection of novelty is harder than the detection of relevant information for a topic. It was found that the detection of novel information within opinion topics was similar to the detection of novel information within event topics. It was also found that the inclusion of training data did not help the overall performance of the novelty detection systems in any year. It was observed that many approaches were applied from other research areas such as the filtering track in TREC and in topic tracking task in TDT to solve the problem of finding new documents within a list of relevant documents by creating systems that performed better than the baseline system however it was noted that novelty detection is not a solved area and remains a hard problem [SH04].

3.5 ImportanceValue Measure

In this section we introduce a new algorithm which we developed for TREC Novelty 2004 based on a traditional information retrieval similarity approach *tf*idf* described in Section 3.14 and word count measures described in Section 3.3. The main aim within novelty detection is to reduce the amount of redundant data that is displayed to the user, thereby inevitably increasing the reader's knowledge in his/her topic of interest. We assume that a user actively gains knowledge on a subject as he/she reads. As a result our approach to the detection of novel information within a sentence compares the current sentence(s_c) to the set of previously seen sentences already calculated as novel and presented to the user. Our algorithm attempts to model our belief that new information contained within a sentence is also important information that a user finds useful to increasing his/her knowledge on a specific topic. We can determine the importance of a word by calculating the frequency with which

it has occurred both within the current sentence (i.e term frequency (tf)) and also by calculating the frequency with which it has occurred over the collection of sentences that the user has already seen to date the "visible document frequency". Words with high term frequencies (tf) and high inverse document frequencies (idf) are most likely valuable or important in providing new and valuable information about a topic. Prior to the implementation of the algorithm the following initial preprocessing steps were carried out. For each sentence within the collection, frequently occurring words or stopwords, such as *and*, *or*, *the*, that offer no valuable information to the reader are identified via a stopword list and removed. Words occurring within all sentences are then put into a word weight matrix which increments a value of one upon the existence of a word within a sentence.

A novelty value is then determined for each sentence within the list of relevant sentences by implementing the ImportanceValue measure, defined as:

$$IV_{s_c} = \left(\sum_{i=1}^{n} tf_{new_{w_i}} \cdot \sum_{i=1}^{n} idf_{new_{w_i}} \right) \cdot \frac{1}{N} \tag{3.14}$$

Where the following notation is defined.

- s_c represents the current sentence under investigation
- new_w represents a new word (i.e. this word has not appeared in *any* sentence seen to this point)
- tf_{new_w} represents the term frequency (tf) of the new word in the current sentence
- idf_{new_w} represents the inverse document frequency of the new word (The reader is referred to Chapter 1 for more information on inverse document frequency)
- N represents the total number of words within the current sentence s_c

- IV_{s_c} represents the ImportanceValue Score of the current sentence s_c (i.e Novelty Score)

The ImportanceValue algorithm is an incremental process and is displayed in Figure 3.2. We will describe a typical walk through the algorithm. Intuitively the initial sentence in a list is always novel see Section 3.1.2. All words within this sentence are considered new and placed within the history set. Thereafter for each sentence s_c in an ordered list of known relevant sentences, we first calculate the number of new words new_w that occur in that sentence by comparing it against the accumulated history set of all the words, which have been encountered in all the novel sentences up to this point. Secondly the *ImportanceValue* states that for all new words new_w within the current sentence we determine the product of both the sum of their term frequencies $\sum_{i=1}^{n} tf_{new_{w_i}}$ and the sum of their inverse document frequencies with respect to the collection of novel sentences already identified $\sum_{i=1}^{n} idf_{new_{w_i}}$. The product is normalised with respect to the length or number of words within the current sentence. The score or novelty value is assigned to the current sentence s_c.

Finally it is necessary to compare the sentence's novelty score against a predefined static threshold θ to determine whether the current sentence contains new or redundant information. If the score for the current sentence s_c is above the predefined threshold, the sentence s_c is added to the list of novel sentences to be displayed to the user. The resulting set of new words new_w from the current sentence are added to the accumulating history set which contains all the words from all the previously seen sentences. This process continues until the entire original collection of relevant documents have been assessed for their novelty value.

3.5.1 Determining Threshold values

Within novelty detection the threshold value, from henceforth known as θ, is necessary to distinguish novel sentences from redundant sentences. Within a

real world application of a novelty detection system the different tolerance levels for the detection of novel information may vary. Users wishing to receive as much information about a topic with as little redundant information as possible in an effort to save time and increase their knowledge simultaneously will require a threshold that highly discriminates redundant data. This is achieved by increasing the threshold value θ thereby reducing the number of sentences that will be returned to the user as novel. Contrary to this however is the case where a user wishes to obtain the details of a particular fact or that doesn't mind viewing redundant or overlapping information. This case would require a decrease in the threshold value θ see Figure 3.2.

We implemented the ImportanceValue measure on the AQUAINT collection of text news data from both the 2003 and 2004 TREC novelty tracks [SH03, SH04]. The threshold value θ that determined the level of novelty detection to be applied to the relevant list for the 2004 data collection was estimated using the 2003 novelty track data collection. A sentence that is assigned a novel score, higher than a predefined threshold θ (set to different values for different collections), is considered a novel sentence. Novelty is determined on a single pass of the results list using a static threshold which was set on the training data. A sentence assigned a novel score, higher than a predefined threshold, is considered a novel sentence.

3.6 Experiments

TREC Novelty provides a common set of guidelines and evaluation measures to allow research groups to test and evaluate the performance of their individual novelty detection systems. This common set of evaluation measures (see Section 3.4.1) allows comparison across different systems. Within the novelty track successful novelty detection approaches are expected to beat the baseline novelty Fscore which returns all relevant documents as novel. We participated in

TREC Novelty in 2004. This participation enabled us to conduct a comparative analysis between our approach and other approaches taken to novelty detection while at the same time enabling us to examine the performance of our novelty detection models against the baseline when implemented on the AQUAINT collection[1] using the common guidelines. The Fscores of all approaches taken by us to the detection of novelty in 2004 is displayed in Figure 3.3[2],[3].

We evaluated the performance of the ImportanceValue measure using two different threshold values (Section 3.5). We also investigated a system called "UniqueHistory" which determined the novel scores of a sentence by calculating the number of new words that occurred in the sentence word set against an accumulating list of all new words that were encountered to this point (for a particular topic). If the number of new words exceeds a particular threshold then the sentence was considered novel which in our runs was defined as three. This is a crude way to determine novelty but as the results show it is a method which gives comparable results. We submitted a total of four runs (see Table 3.2). The Baseline run (cdvp4NSnoH4) used the UniqueHistory measure however we did not keep an accumulated history set of all the previous sentences.

Table 3.3 shows the Fscores of the top performing novelty detection systems for Task 2. The ImportanceValue measure algorithm was the highest performing novelty detection system run of 2004 achieving an Fscore of 0.622 [SH04], with a threshold value of 1.5. From this table it can be seen that the ImportanceValue measure not only out-performed the other systems observed through the Fscore values but that the accuracy in finding new sentences is also quite high, which is evident from the precision value.

A key aspect of utilizing our "ImportanceValue" measure is the threshold θ

[1]The Aquaint collection is a corpus of approximately 1,033,000 documents or 3GB of English news text.

[2]This figure was taken from the overview slide of TREC novelty 2004 presented by Ian Soboroff

[3]The CDVP novelty approaches are highlighted by arrows in Figure 3.3

Task2 Novelty Measure	Run	Precision	Recall	Fscore
Average(mean)				0.576
ImportanceValue > 1.5	cdvp4NTerFr1	0.49	0.90	0.622
ImportanceValue > 3.5	cdvp4NTerFr3	0.51	0.83	0.616
UniqueHistory > 3	cdvp4UnHis3	0.50	0.84	0.615
CDVPBaseline	cdvp4NSnoH4	0.38	0.49	0.383

Table 3.2: Description of all our runs submitted to Task 2 of Novelty Track 2004

Group	Precision	Recall	Fscore
Return All Documents(Baseline)			0.577
Average Fscore			0.576
Meiji Uni.	0.48	0.93	0.619
Uni. of Mass	0.47	0.95	0.618
ImportanceValue	0.49	0.90	0.622

Table 3.3: The Fscore of runs in 2004

Group	Precision	Recall	Fscore
Return All Documents(Baseline)			0.774
Average Fscore			0.689
ImportanceValue Measure	0.73	0.94	0.808

Table 3.4: The Fscores achieved in 2003

above which we assume a sentence to be novel. The initial threshold values were determined on a subset of documents manually extracted from the 2003 novelty data collection. Subsequent to the novelty track, experiments were extended and the threshold values were optimised. We examined a range of threshold values using the 2004 data, as shown in Figure 3.4. Optimizing the threshold did not provide a significant improvement (Fscore 0.623) over our previous official TREC novelty run.

Although we had not participated in TREC2003, we carried out the same procedure on that data with an optimised threshold for 2003 (see Figure 3.5) yielding an Fscore of 0.808. In 2003 there were forty five runs submitted to the Novelty task. This Fscore would have placed us sixth highest among novelty runs showing that the ImportanceValue algorithm is a robust technique to detect novelty on different data collections see Table 3.4.

The Fscore from our runs on the 2003 data at 0.808 is larger than that obtained on the 2004 data with an Fscore of 0.622. Although the data for 2003 and 2004 came from essentially the same resource this variation in thresholds is certainly not unexpected. It has been shown in other TREC tracks, such as TRECVid that even though data may come from the same source two years in succession, optimization for different years produces different best parameter values and different best performances.

There are a number of possible reasons for this including the fact that topics for each of the years are different, with the topics for 2004 proving more difficult overall. The average Fscore on all topics for 2003 was 0.731 and for 2004 it was

0.597. The average precision for each topic for 2003 was 0.652 whereas for 2004 it was 0.46.

3.7 Summary

In this Chapter we introduced the concept of novelty detection in information retrieval. We have seen how there are three types of novelty detection namely the detection of novel events over an entire collection, the detection of new stories about a known topic and the detection of novel information from within a retrieved results set for any user specific topic. The latter is the research focus of this thesis. We outlined six assumptions that are made in order to avoid ambiguity during the identification of a "novel" document. We looked at the novelty track in the annual TREC conference, the collections, the topics, the evaluation procedures and finally two of the best performing approaches from each year. Finally we introduced our novelty detection algorithm, the "ImportanceValue" measure, which was developed for the novelty track in 2004. We looked at its performance over both the 2003 and 2004 novelty data collections using the common evaluation measures.

In the following chapter we introduce novelty detection on visual broadcast news. As we have see in Chapter 2 video is composed of a multiple components including audio, visual and semantic layers. It was shown that the "ImportanceValue" measure was a robust technique in the detection of novel data from textual new data and as a result we apply the algorithm to detect novel data from the textual component of digital video.

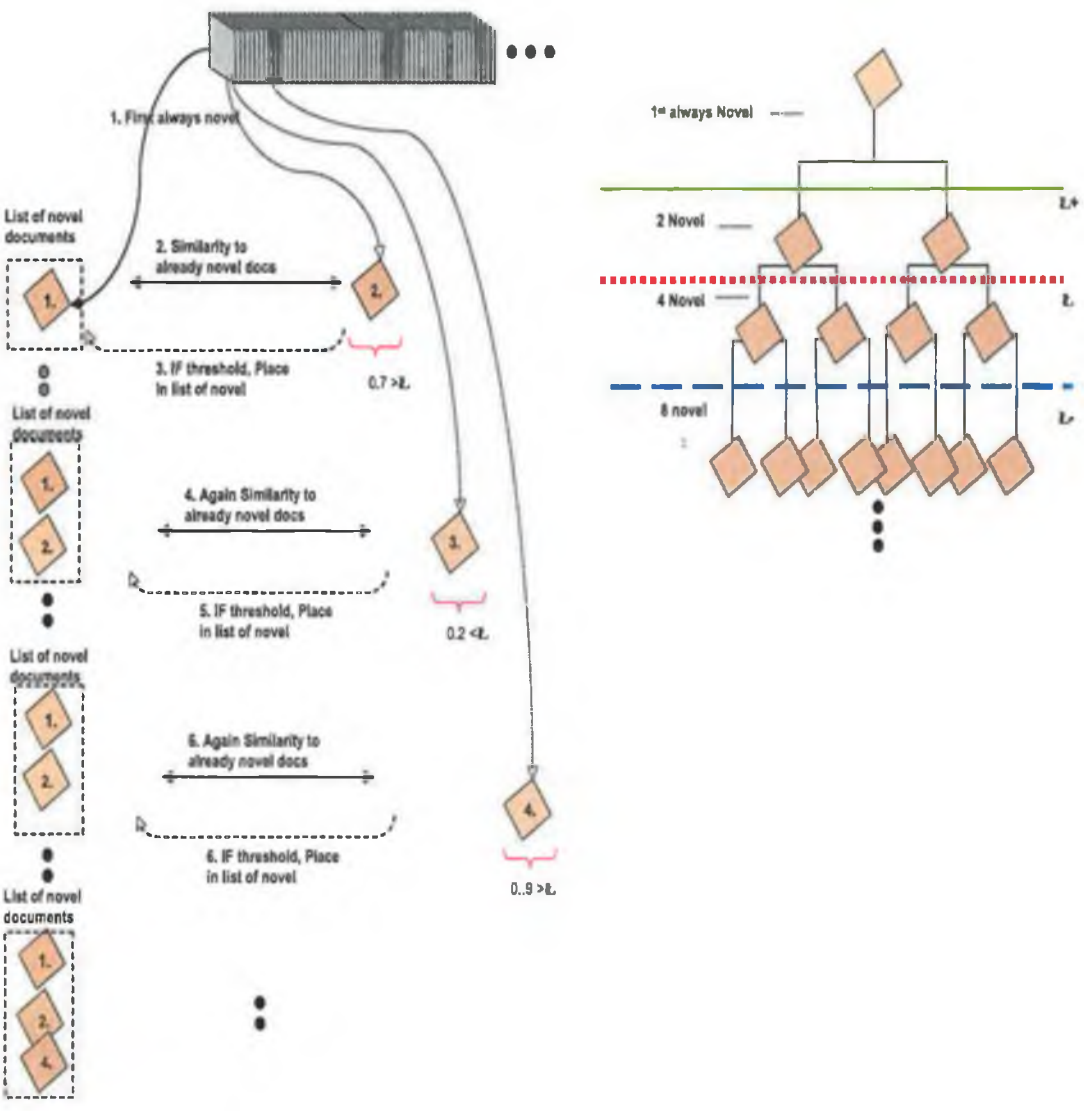

Figure 3.2: Novelty detection architecture using the ImportanceValue Algorithm. The higher the threshold value the fewer number of documents will be considered novel

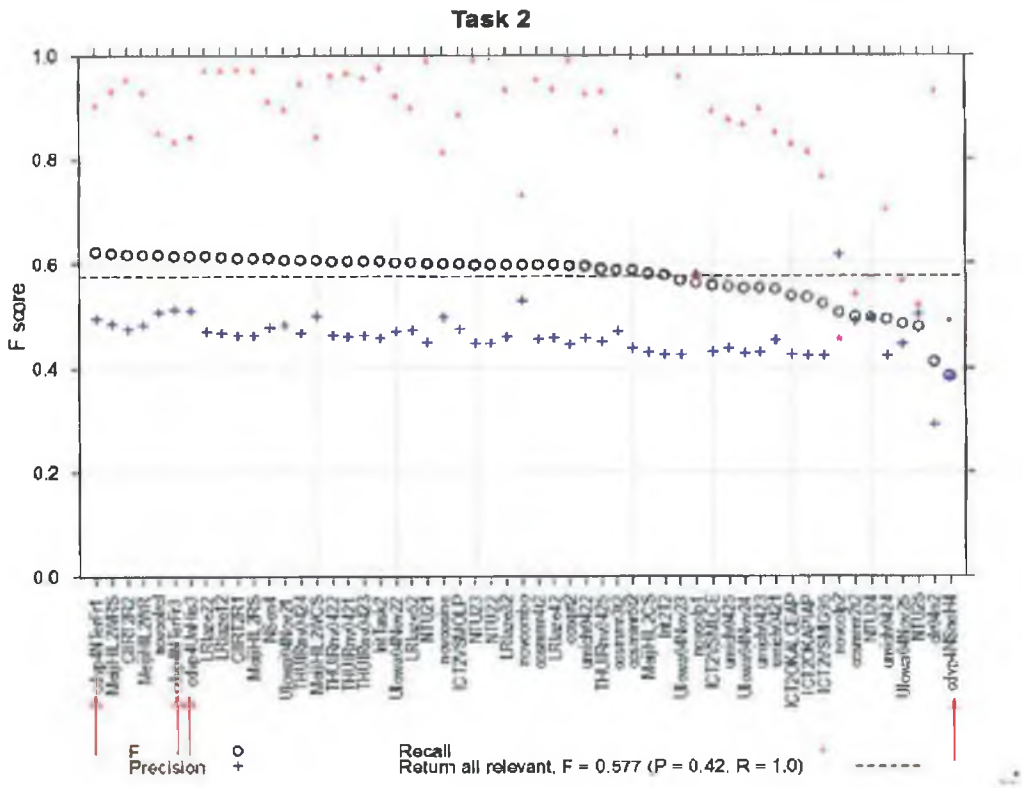

Figure 3.3: Novelty runs for TREC Novelty 2004 data

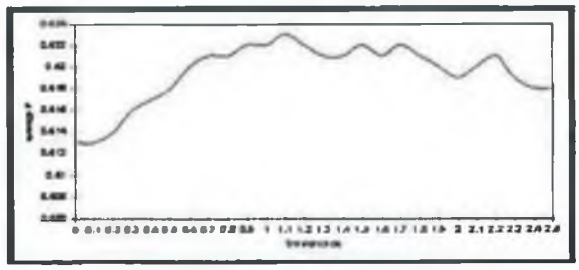

Figure 3.4: "ImportanceValue" Fscores vs. threshold on 2004 data

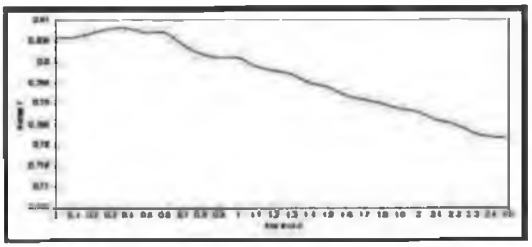

Figure 3.5: "ImportanceValue" Fscores vs. threshold on 2003 data

Chapter 4

Novelty Detection in the Context of Video

In this Chapter we look at Novelty detection in the context of content based video retrieval. We will discuss the detection of novel information from within a search output for any user specific topic within the video domain. This will motivate the need for novelty detection in content based video collections and in a particular TV news broadcast collections. We will then look at the challenges and complex issues that arise when designing models for video, such as overall video structure or the multiple modalities associated with a video sequence that can be extracted and offering valuable information. We will then discuss the considerations that must be taken into account when designing an overall novelty detection model for video. Finally we will introduce a novelty detection model designed to accurately identify novel shots from a results list, which we use in experiments reported later.

4.1 Novelty Detection in Content Based Video Retrieval

One of the major challenges in the information retrieval community is the accurate retrieval of information across different media. Within the text retrieval domain [AAB+02] much research has been conducted into retrieval models that aid retrieval and reduce a user's cognitive load. These include the language modeling approach, as well as models for web search, filtering, topic detection and tracking, classification, novelty detection, summarisation and question answering. These attempt to improve the users' overall searching experience when looking for his or her desired information and have resulted in many sophisticated, mature and well documented approaches being implemented and evaluated within the text domain.

To date most research within the video information retrieval community has concentrated on improving search and browsing facilities although recently there has been activity in research that explores content based techniques for the automatic summarisation of video collections. As of yet these techniques have only been realized on specific knowledge domains such as news or sports. At the time of writing, little if any, research has been carried out into question answering or the detection of novel information from within a retrieved results set of documents for any user specific topic within the content based video domain. There is no particular reason for this apart from the immaturity of the video information retrieval field.

4.1.1 The Motivation for Novelty Detection in Video Retrieval

A typical broadcast TV news program is usually a very rich source of information on a variety of diverse news topics. These programmes record the evolution of a news story in time and contain valuable information for creating documentaries. However it is also rife with repetition as news broadcasters frequently

use previously seen video footage on a continuous basis, either in an attempt to remind the viewer of a past story, or as a headline to introduce what is about to be presented within the broadcast, or indeed as a summary of the news programme. Often when a story breaks, broadcasters may not have a reporter in the area and will reuse old video sequences taken from an archive which were broadcast previously about a similar event. If a collection contains different news programmes from different broadcasters, many stories describing the exact same information with perhaps a slight variation of commentary or imagery may be repeated across broadcasters. A typical video retrieval system will return the repetitive video sequences including those containing exactly the same video footage or graphics which are contained within a collection, as part of the result list in response to a particular topic. This redundancy degrades a users' overall search experience with a system as he/she is required to sift through the superfluous information in search of previously unseen data. Due to the growth in the television news sector it is becoming necessary to develop "intelligent" methods that determine the novelty value of the information presented. As a result novelty detection systems could have very real applications in the area of multimedia information retrieval and particularly in the genre of broadcast TV news.

Here we seek to organise broadcast news search outputs based on the degree of "newness" to the search topic rather than ranking by degree of relevance. Novelty detection techniques have already been applied successfully to the text domain to combat such problems [Har02, SH03, SH04] as shown in a previous Chapter of this thesis. In the following sections of this Chapter we will discuss a model designed to detect novelty within a video collection.

4.2 Considerations in Designing A Novelty Detection Model

In Chapter 3 we defined the notion of novelty detection within the text domain as the identification of novel documents from within a results list that was returned for a user specific topic. This definition for novelty detection holds true within the video domain, however instead of dealing with documents or sentences as in the text domain, novelty detection within the video domain deals with video shots. As a result a novelty detection model within video is concerned with identifying novel shots from within a list of results returned to the user for a specific topic, thus bringing some interesting and challenging issues to the fore.

The analysis of video is quite a complex challenge. Multimedia is far more difficult to manipulate than text, mainly due to the fact that, unlike text (where we can attempt to deduce the semantic meaning through words), we have no standard way of extracting the semantic meaning from an image, to say nothing of doing the same from a video clip ! Text spoken during a shot is not a sufficient method of assessing a shot's novelty value as visual content is not aligned with spoken content, this is clearly evident during the commentary of a sports event. This has resulted in little research being carried out into alternative methods of search and retrieval within the video domain.

4.2.1 Representation of video

As discussed in Chapter 2, a video sequence can be broken up into a hierarchical structure with the frame considered the most basic component. However to analyse a video based on frames alone can become computationally expensive and, in many cases, redundant as frames from specific shots are very similar and if evaluated separately will contain much visual redundancy. Scenes contain a number of shots grouped into a logical combination depicting a story/event.

However scene detection is not accurate over most genres (with television news being the exception) and scenes are generally too broad to accurately represent information content. The shot is widely accepted within the video information retrieval community as the basic unit of retrieval for video-based retrieval systems. Shots can be detected automatically with reasonable and acceptable accuracy and are small enough to represent the information contained within the video document. In order to increase efficiency, computation is carried out on one representative frame, known as the "keyframe", for each individual shot.

A novelty detection model designed for the video domain will use shots as the base unit of manipulation, with the multiple modalities of video aligned at the shot level where the demarcation is clear. One keyframe within each shot is extracted as a representative image.

4.2.2 Novelty detection as duplicate detection

Shot detection and segmentation techniques do not currently achieve one hundred percent accuracy and thus a video sequence may have several shots, with the keyframes for each shot differing only slightly due to a different camera angle or action such as zooming in or out. It must also be noted that the keyframes may be visually very similar but not necessarily identical and this is clearly evident from Figure 5.2.2 which shows four very similar looking shots that have been returned within the results list for TRECVid Topic 125. In a standard comparison measure, such as the Manhattan Distance Measure, two shots are identical if their dis-similarity values are zero. However when "shot36_186", the query shot in this example, is visually compared for colour using the Manhattan distance against each of the other shots sequentially, it is evident that the shots are not necessarily duplicates according to the Manhattan Distance Similarity measure values. This example highlights the fact that simple detection of duplicate shots is not sufficient to remove redundant and uninformative information from a results set.

Current Shot id	Shot id	Manhattan distance
shot36_186	shot17_99	6.7082
shot36_186	shot14_91	6.4031
shot36_186	shot16_76	6.6332
shot36_186	shot36_186	0.0(duplicate shot)

Figure 4.1: Example of four very similar shots namely shot17_99, shot14_91, shot16_76 and shot36_186 respectively

4.2.3 Evolution of Stories

In Chapter 3 we looked at the chronological ordering of documents within the novelty detection track of the text domain. Documents corresponding to TV news stories were ordered chronologically due the fact that, in theory, as a story evolves, earlier stories on a particular topic will contain a lot of unknown or unseen information while stories coming later will not contain as much new information on the particular topic. This theory remains true within the video domain. As a result shot timestamps are a very important aspect/attribute to consider when assessing the novelty of a shot and should be incorporated into the overall design of a novelty detection model. In this thesis, our model for novelty orders shots chronologically with the oldest shot appearing first or highest in the results list.

4.2.4 Human perception of images and interpretation of novelty

There have been many studies on why and how images are much more difficult to index than text. The subjectivity inherent in pinning down what is depicted in a picture has been studied in depth in an attempt to develop some generic

method of indexing audio-visual libraries [Sha86, Lay94, Ens95]. What one person might consider important in an image may differ from what another person would consider important within the same image. Consider for example Figure 4.2 containing three keyframes of a hockey game. Is the second keyframe

Figure 4.2: Example images of a hockey game

novel or is it providing no new information when compared to the first keyframe? The location of a hockey player may be noticed by some assessors for example, but not by others [Sha86]. Unlike the factual information depicted in text format ("Columbus discovered America in 1492") it is difficult to determine whether or not you have seen this particular image of a hockey game before without going back and checking it. It is therefore necessary, when assessing a shot's novelty value against one or many shots in a set, to perform the task on a shot to shot basis, where the shot under investigation is compared to each of the shot separately for each shot within the set of novel shots found to this point, rather than against an entire set, where the shot is compared to the collective characteristics of all shots within the novel set on a first pass. It is also necessary to record a decision about its novelty value against a particular shot immediately, before continuing to the next shot in the set. The overall determination of a particular shot's novelty will then be based the accumulation of the shot's novelty values against all shots in the set. If the resulting novelty value is of a sufficient level, then the shot would be considered novel otherwise it would be considered redundant. As a result, novelty detection within the video domain is far more difficult to determine than within the text domain. This subjectiveness in perception has also led to a subjectiveness within the ground truth data with different assessors having different opinions on a shot's novelty value, based on what they perceive important in the shot. This is described in

more detail in Chapter 5.

As the novelty detection from within a results list in the video domain is a completely new area of research it is important to observe human interpretation of and interaction with the task, and develop the novelty model as accurately as possible based on this gathered information.

4.2.5 Categorisation of queries

Currently within the video retrieval community the classification of user queries or topics into specific predefined classes that contain queries of a similar type is a research topic that has been gathering a lot of attention recently and has been successfully implemented by CMU [yYH04]. These possible query classifications include:

- People: All queries relating to people including the actual person in question or a physical action performed by that person. An example of such a query would be Topic 133 "Find shots of Saddam Hussein". Topics belonging to this category from TRECVid2004 include Topics 128, 134, 135 137,and 144.

- Specific Object: All queries relating to a uniquely named object or entity, distinguishing the object in question from all other objects of the same type. An example of such a query would be Topic 129 "Find shots zooming in on the U.S Capitol dome". As it happens this is the only topic in TRECVid2004 that belongs to the Specific Object category.

- General Object: All queries relating to certain types of objects rather than one specific object. An example of such a query would be Topic 140 "Find shots of one or more bicycles rolling along ". Other topics from TRECVid2004 belonging to this category include Topics 132, 139, 141, and 143, 145.

- Sports: All queries relating to a sports event. An example topic from the TRECVid 2004 collection is Topic 130 "Find shots of a hockey rink with at least one of the nets fully visible from some point of view". Other topics in the category include Topic 136 and Topic 142

- Other (Scenes): All queries depicting multiple types of objects and their surrounding environments or spaces. An example topic from the TRECVid 2004 collection includes Topic 126 "Find shots of one or more buildings with flood waters around them". Other topics in the category include Topics 125, 127,131, 138 , 147 and 148.

The topics and their associated categories are described in more detail in Appendix 7.3.

4.2.6 Using Multi-modal resources

As described in Chapter 2, video is composed of many modalities, including text, low-level feature evidences and higher level semantic evidences, all of which are valuable resources that can be utilised to determine a shot's novelty value when compared to a previously seen shot. We believe a novelty detection model within the video domain should be broken up into several novelty components capable of incorporating and extracting information from these invaluable resources individually to assess the overall novelty of a shot. The main components of a novelty detection model for video are listed below and described in more detail in section 4.3:

- Low-level features novelty component: The model will need to be able to assess the novelty value of a shot when compared to another shot based on individual low-level features. The model will further need to be able to combine the novelty values for each of the various low level features including colour, edge and texture to achieve an overall novelty value based on all features for the shot. Features vary in the properties and

ranges of values and dimensions and so combining them is not a trivial process.

- Text novelty component: There are text portions in the form of automatic speech recognition (ASR) transcripts available for all shots within the broadcast TV news video collection we are using, corresponding to the dialogue that was spoken.

- Automatic high level features novelty component: Each shot has associated with it a set of high level features, such as anchor person, commercial, face and person to name a few.

- Manually annotated concepts novelty component: The novelty model will need to incorporate the manually annotated information for each shot.

Novelty detection models have been successfully developed within the text domain (see Chapter 3) and intuitively, it should possible to adapt these models to assess the novelty value of a shot using its associated text portion only. However not all shots in a video collection will have an associated text portion and in such cases the identification of a shot's novelty value will rely on the visual evidences associated with that shot. This issue will be further discussed in Section 4.3.1.

The model should be also be capable of combining or unifying the text and low-level feature components, text and automatic high-level concept components and, finally, the text and manually annotated concept components in an attempt to further assess a shot's novelty value.

In the next section we will describe the novelty detection model that was designed for the detection of novel shots from within a results list of shots relevant to a specific user defined topic.

4.3 A Model for Video Novelty Detection

The novelty detection model was designed to closely mimic a human being's interaction with and interpretation of the novelty detection task, given a topic and a chronologically ordered list of relevant results to that topic. It was observed that the assessment of a shot's novelty value was performed on a shot to shot basis for all shots in a set rather than on a shot to set of shots basis. So given a shot s_i and a list of previously seen and novel shots so far $L = snovel_1, snovel_2, snovel_3, ..snovel_n$, s_i is against $snovel_1$ to determine its novelty value, then it is compared against $snovel_2$, then $snovel_3$ and then each in turn until s_1 has been compared against $snovel_n$ for a novelty value. The shot s_1 was considered redundant if the contents of the shot were previously seen by a shot in the list of previously seen shots L. This technique was not unexpected as a similar trend occurs in novelty detection in text documents [ZCM02] though it is more obvious in the video domain due to the perceptual level with which video is assessed or viewed as discussed earlier in section 4.2.4.

A Generic Algorithm for Novelty Detection

In this section we describe the generic algorithm for the detection of a novel shot using any of the video resources as a means of detecting a shot's novelty value.

- Consider a list of shots returned to the searcher for a specific topic.
- The first shot in this list is always novel as per assumption 6 in Chapter 3 section 2 which assumes that a user knows nothing about the topic at the initial shot.
- The searcher views each of the subsequent shots in sequential order.
- The shot must contain a certain level of novel information when compared to the set of previously seen shots in the novel set in order to be classified as a novel shot.

- If the shot is classified as a novel shot, it will be added to the novel set (the set of previously seen shots).
- The process continues for each subsequent shot until all shots have been classified as either novel or redundant.

The main objective of the novelty detection model is to maximise the inclusion of novel samples while at the same time minimising the inclusion of known or previously seen (redundant) samples. It is independent of the content retrieval methodology. The novelty detection designed and described in this thesis for the video domain consists of the four main components including text, low-level feature, automatic concept and manual concept components as outlined in section 4.2.6 each of which are described in the next section.

4.3.1 Novelty Model:- Text Component

During human assessment, the audio associated with the video shots was removed because of the additional significant complexity it introduced to the novelty detection task. As a result, assessors made their judgement on a shot's novelty based solely on the visual evidence presented to them in each shot. However it has been consistently proven that text is a very valuable resource in traditional video retrieval systems [BCG+03, BCG+04, Hom05] and so we developed two novelty models designed to accurately identify novel shots within a results set given only the textual data associated with each shot. The first model was designed to assess the novelty of a shot by comparing the shot to the entire set of previously seen documents, while the second model was designed to assess a shot novelty value similar to the interaction of a human's assessment of novelty, namely on a shot to shot basis for all shots within the set of previously seen documents.

In order to do this, text in the form of automatic speech recognition (ASR) transcripts, supplied by LIMSI [JGA02] and provided by TRECVID, was segmented and aligned with each individual shot in the collection.

In Chapter 3 (section 3.5), we looked at a successful model, the Importance-Value Measure, equation 4.1, designed for novelty detection within the text domain. This model was implemented and evaluated on the TREC Novelty AQUAINT collection (text news data) [SH04] and produced good results. For the purpose of the novelty track experiments, sentences were considered as documents. There are parallels between sentences in documents for text novelty detection, and shots in video news stories[1] and we inherit this characteristic from the TREC track to allow comparability. Consequently the Importance-Value measure was adapted and employed on the shot textual portions to detect novelty among shots within the video domain.

The model performs novelty detection in the video domain in a manner very similar to that of detecting novel sentences in the text domain. The algorithm's structure is summarised as follows:

- Given a list of relevant shots, the first shot is novel and the algorithm iteratively takes as input the next shot on the relevant list.
- Each shot is analysed for novel words against the set of previously seen and declared novel shots.
- For each unique word found within the shot the term frequencies and inverse document frequencies are calculated and provided as input to the ImportanceValue measure resulting in a novelty score.
- If a shot achieves a novelty score above a certain predefined novelty threshold value θ it is considered a novel shot and is consequently added to the accumulative novel set.
- Otherwise the shot is considered redundant
- The process starts again with the next shot on the list.

[1]The average shot length for TRECVID 2004 shots was 12 terms not including stopwords. This is very similar to the average sentence length consisting of approximately 15-20 terms not including stopwords

However, as mentioned earlier, some shots do not have an associated text portion and in such cases the original ImportanceValue measure would fail as it is principally looking for unique or novel words. It is not possible to make a decision on a shot's novelty score based solely on text if no textual data exists. To this end the ImportanceValue measure has been adapted to allow for such incidences by making a non textual shot novel by default. This allows other modalities to influences the novelty value of a shot using the combination of multiple modalities.

If $N > 0$ then

$$Nov_{s_c} = \left(\sum_{i=1}^{n} tf_{new_{w_i}} \cdot \sum_{i=1}^{n} idf_{new_{w_i}} \right) \cdot \frac{1}{N} \tag{4.1}$$

otherwise $s_c = novel$ by default.

If $Nov_{(s_c)} > threshold$ then $s_c = novel$
otherwise $s_c = redundant$

where the following notation is defined.

- s_c represents the current sentence under investigation
- new_w represents a new word i.e this word has not appeared in *any* sentence seen to this point
- tf_{new_w} represents the term frequency (tf) of the new word in the current sentence
- idf_{new_w} represents the inverse document frequency of the new word (the reader is referred to Chapter 1 for more information on inverse document frequency)
- N represents the total number of words within the current sentence s_c

The second model represents the behaviours noted during a human assessment of the novelty task. It is involves a further adaption of the ImportanceValue

model (see equation 4.2). In this equation a shots novelty score is initially determined by comparing the shot against each shot within a set of predetermined novel shots. The minimum novelty score over all shots in the novel set is extracted (see equation 4.3). A shot is considered novel if the minimum novelty score is above a certain predefined novelty threshold value and is added to the accumulative novel set. Otherwise the shot is considered redundant and the processes continues until all shots in the results list have been classified. Once again non textual shots are assumed novel by default.

If $N > 0$ then

$$Score(s_c, s_j) = \left(\sum_{i=1}^{n} tf_{new_{w_i}} \cdot \sum_{i=1}^{n} idf_{new_{w_i}} \right) \cdot \frac{1}{N} \tag{4.2}$$

$$Nov_{s_c} = \min_{j=1}^{m} (Score(s_c, s_j)) \tag{4.3}$$

otherwise $s_c = novel$ by default.

If $Nov_{(}s_c) > threshold$ then $s_c = novel$
otherwise $s_c = redundant$

In the next section we will look at the novelty model designed to assess a shot's novelty values when compared to another shot based solely on the low-level features that are contained within both shots.

4.3.2 Novelty Model:- Low Level Features Component

The visual novelty detection model developed for the visual aspects of the video is very similar to the text novelty model which was described in equation 4.2 section 4.3.1, however the method of shot comparison is different.

Initially a shot's similarity score is determined against each of the individual shots within a set of predetermined novel shots (see equation 4.4). This is achieved by first, calculating the similarity scores, using the Manhattan Distance Measure, for each of the available features $F_1..F_k$ of the shots being compared independently. The similarity scores obtained for each of the features

are then linearly combined to obtain one novelty score for the shot in question. This process is continued until the shot in question has been compared against all shots in the novel set.

The minimum novelty score achieved, for the particular shot against all shots $S_1..S_m$ in the novel set, is then extracted (see equation 4.5).

A shot is considered novel if the minimum novelty score is above a certain predefined novelty threshold value θ and is added to the accumulative novel set. Otherwise the shot is considered redundant and the process continues by taking the next shot on the list until all shots have been classified.

$$Score(s_c, s_j) = \sum_{i=1}^{k} Sim((F_i(S_c)), (F_i(S_j))) \tag{4.4}$$

where $F_i = i^{th}$ feature of the shot

$$Nov_{s_c} = \min_{j=1}^{m}(Score(s_c, s_j)) \tag{4.5}$$

If $Nov_{s_c} > threshold$ then $s_c = novel$
otherwise $s_c = redundant$

The shot comparison method used, the Manhattan distance $Sim(S_c, S_j)$ described in section 4.3.2, is a dissimilarity measure, so the smaller the value the more similar the shots actually are.

Visual features are often represented as histograms which clearly depict the features' distribution across a feature "space" or set of possible values. These histograms can be represented as a vector describing the visual content. This allows accurate similarity comparisons defined in terms of the distance between the vector representations, to be performed between the shots. Feature vectors can also be normalised prior to calculating similarity distances allowing the accurate combination of many varying features with varying dimensions. This is discussed in more detail in section 4.3.2.

The visual features that are used to model the low level visual component of the novelty model are described below and include two MPEG-7 descriptors,

the MPEG-7 colour structure which has been successfully incorporated into video retrieval systems designed for TRECVid [PHO+02] and the MPEG-7 edge histogram, as well as three denoted features including colour, edge and texture which were supplied by CMU to the participants of TRECVID2004. Each of these are discussed in more detail in the next section.

Low-level Visual Features Evidences

In this section we describe each of the low level features that were used to help identify novel information from a video keyframe.

MPEG-7 Colour Structure: evidences are defined within the MPEG-7 (XML-like standard) standard [Com02]. This is a histogram-like feature that describes the colour contained within an image while also providing information about the structure of this colour content in the image. Colour is represented using the HMMD colourspace which defines five dimensions - Hue, Max(max of R,G,B triplet), Min(min of R,G,B triplet), Diff($Max - Min$) and $Sum(\frac{Max+Min}{2})$. The colour structure is calculated using an 8x8 pixel square window that slides over the the entire image. It increments the counts for each colour encountered in the window as it slides over the image.

MPEG-7 Edge Detection: evidences are defined within the MPEG-7 (XML-like standard) standard [Com02] describing the edges within an image using an edge orientation histogram. It defines an image by using a 4x4 grid (16 rectangular regions) and identifies four directional edges (horizontal, vertical, 45 degree diagonal, 135 degree diagonal) and one non-directional. The histogram bins are normalised with respect to the number of pixels found in the image under investigation.

HSV Colour: The image is divided into a 5x5 grid. The colour evidences represented in histogram form as denoted by CMU are extracted using the HSV

colourspace. HSV is a perceptual-based model separating the colour dimensions into Hue, Saturation and Value (brightness) commonly used within video and image retrieval systems [WK96].

Canny Edge: The edge feature evidences, represented in histogram form, were extracted using the Canny edge detector [Can86] on each keyframe which was split into a 5x5 grid.

Gabor Texture: The texture evidences, represented in the form of a histogram, were extracted using Gabor filters. A Gabor filter is a modulated product of Gaussian envelops and sinusoidal signals and is defined in equation 4.6.

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp^{(-1/2(\frac{x^2}{\sigma_y^2}+\frac{y^2}{\sigma_y^2}))} \exp^{jWx2\pi} \tag{4.6}$$

where σ_x and σ_y represents the standard deviations of the Gaussian envelop and W represents modulation frequency. By applying various scales (standard deviations of the Gaussian envelopes) and orientations, texture evidences can be extracted from an image. Each image is converted to greyscale and divided into a 5x5 grid. Six orientated gabor filters are then applied to these greyscale images.

We take a black box approach to these features for integration into the novelty detection model in that we do not experiment with them and try to optimise parameter settings. We assume that each feature has a list of shots with an associated confidence value. This value represents the feature detector's confidence in the evidence for that feature being present within the shot.

Visual Similarity Measure metrics

Although the modeling methods for visual features such as colour, edge, texture within an image are semanitcally different, the detection of features is usually represented using either a vector or histogram. Many methods have been suggested for comparing the vector or histogram representations of images. These include the Histogram representation described in [SB91], the statistic proposed by [PHB97], Relative Entropy as described in [OPH96] and the Jensen-Shannon method described in [Rao82]. The standard measures for the comparison of two vector or histogram representations of an image within the IR community however, are the Minkowski form distance measures, the Manhattan distance and the Euclidean distance.

The Minkowski form distances The comparison of two normalised feature vectors $F(S_1)$ and $F(S_2)$ containing k elements and representing shot S_1 and shot S_2 respectively is usually carried out using some form of the Minkowski distances measures defined as:

$$s_{L_p}(S_1, S_2) = (\sum_{i}^{k} |F_i(S_1) - F_i(S_2)|^p)^{\frac{1}{p}} \tag{4.7}$$

where $F_i = i^{th}$ feature component in the normalised vector

When $p = 1$ we have the Manhattan distance (L_1 norm or city block distance). Given any two shots S_1 and S_2 the dissimilarity between them can be obtained as the sum of absolute difference between each pair of components $F_i(S_1)$ $F_iS_2)$. Consider the example in Table 4.3.2 where we have 3 bins which represent different and non-overlapping colour ranges while the histogram's values contain the number of pixels in that particular range

bins/Components	Histogram(Shot1)	Histogram(Shot2)	Absolute Score
1	125	100	25
2	100	90	10
3	40	80	40
Manhattan			75 dissimilar

Table 4.1: Manhattan Distance Example

- $p = 2$ defines the Euclidean distance (L_2 norm or as-the-crow-flies distance) which derives the similarity between two shots by computing the square root of the sum of the squares of the differences between the corresponding components
- $p = \infty$ defines the maximum distance between vector elements L_∞.

Our novelty model will use the Manhattan distance as the shot to shot comparison approach for all visual feature evidences. This measure was chosen as the Euclidean distance generally preforms worse for low level images such as colour and edge for the TRECVid search task. [BCG+03, BCG+04, Hom05]

Normalising the Visual Feature evidences

Having described the various features that are extracted from the keyframes of each shot within the collection, the overall novelty score for each shot can be calculated using various combinations of these features. Each of these features, however, contain their own unique set of characteristic values. Consider for example Table 4.3.2.

In this example, colour is represented by very high similarity values while texture on the other hand, is represented by small values. A combination of all features using a linear summation approach would be dominated by the colour feature while the texture and edge have little effect on the overall score. In

Feature	Similarity Score
Edge	150
Texture	46
Colour	1100
Linear Summation	1296
Averaging of features	432

Table 4.2: Feature Combination

order to ensure equal emphasis of each feature (and hence equal emphasis on each novelty value) normalisation is performed. There are a number of normalisation methods by which this may be achieved including a basic averaging of all features where the sum of all features is calculated and then divided by the total number of features represented as can be seen in Table 4.3.2. While this method ignores large differences between features, it fails to alleviate the problem of dominance as high or low values for an individual feature can distort the overall average for the shot.

Another approach that can be employed in an attempt rectify the over dominance of one feature compared to all other features in a shot, is known as histogram normalisation. In this approach normalisation is performed prior to a shot by shot comparison on each histogram representing each individual feature within a shot. The approach works by dividing the count in a bin of the histogram by the total number of counts observed in all bins. The relative counts, overall bins in the normalised histogram, sum to one (or 100 if a percentage scale is used) see Table 4.3.2.

As each feature contains an overall bin summation value of one, it is ensured that no one feature dominates the combined novelty score for each shot. The novelty model in this thesis uses the normalised histogram approach to normalise the features. Once the feature histograms are normalised, visual shot-to-shot comparisons are performed on each feature separately using the Manhattan distance measure. These resulting similarity scores are all within a predefined

bins/Components	Histogram	**Normalised Histogram(bin/count)**
1	125	0.471
2	100	0.377
3	40	0.151
Total **count**	265	0.999

Table 4.3: Histogram Normalisation

range of [0, 2]. These feature similarity values are then combined using linear summation to give an overall similarity value, or novelty score, for the shot.

4.3.3 Novelty Model:- High level/Semantic Concept Component

As discussed in Chapter 2, the detection of semantic information from within a video sequence is very important. The automatic extraction of the semantic meanings from a visual image or video sequence is, however, a highly complex task. In this section we will describe two novelty detection components designed for automatically detected high level features and for manually annotated high level/semantic features. In addition, we will discuss an ontology designed to annotate a video for use in applications such as novelty detection.

Automatically detected high level concepts

The details of methods used for detection of high level/semantic features detections is beyond the scope of this thesis, however Naphade and Smith [NS04] give an overview of the detection approaches that have been undertaken over the last few years. As a result (and in a similar manner to low-level features), we take a black box approach to these features and incorporate them into the novelty detection model directly without any experimentation to try to optimise their settings. We assume that each feature has a list of shots with an

associated confidence value. This value represents the feature detector's confidence in the evidence of that feature existing within the shot. This score is integrated into the MPEG-7 description of a video. The automatically detected semantic features incorporated into the novelty detection model were donated to participants of TRECVid2004 by CMU and include; Face, Anchor, Commercial, Studio setting, Graphics, Weather, Sports, Outdoor, Person, Crowd, Road, Car, Building and Motion.

Research into the most effective feature combinations and corresponding optimal weights for retrieval performance has been carried out by Rong Yan et al. at CMU [yYH04] for each query category mentioned earlier. In his paper he discusses a retrieval model that firstly uses each of the various feature detectors (such as face, anchor and commercial) separately to determine the shot similarity value to a query based on that feature, then combines these multiple evidences of similarity using each feature's associated optimal weights for the specific category of which the query is a member. This paper outlines an effective feature combination and optimal weight for each of the specified categories. These optimal combination weights are used within the development of the novelty detection model for high level features to assess a shot's novelty score.

We adapted the visual novelty detection model described in section 4.3.2 equation 4.4, to incorporate a weighted, linear combination of similarity values for each feature within the shot. The weight chosen for each feature is dependent on the topic under investigation (see equation 4.8).

$$Score(s_c, s_j) = \sum_{i=1}^{k} \lambda_i(c_q) Sim((F_i(S_c))(F_i(S_j))) \tag{4.8}$$

- where $F_i = i^{th}$ feature of the shot
- and $\lambda_i(c_q)$ is the optimal weight for feature F_i when the query is member of category c_q where $q = 1..5$ representing the 5 categories

The overall novelty score for a shot is then calculated by assessing the overall minimal similarity value over all shots $S_1..S_m$ in the novelty set.

$$Nov_{s_c} = \min_{j=1}^{m}(Score(s_c, s_j)) \tag{4.9}$$

If $Nov_{s_c} > threshold$ then $s_c = novel$
otherwise $s_c = redundant$

The shot comparison method $Sim(S_c, S_j)$, uses the Manhattan distance which was described in section 4.3.2. This is a dissimilarity measure, so the smaller the similarity value, indicates the more similar the shots actually are.

Manually annotated semantic concepts

The most common way to index video for content-based retrieval in real-world applications is to use manual annotations of some kind. Automatic feature detectors for video have been developed to accurately identify the presence of a number of specific concepts within a video. Although they work well within narrow domains, for example in soccer, in broader domains such as broadcast TV news where video footage is unpredictable and varied, annotation cannot be accurately achieved. This is because technology has not as yet reached the stage where detectors are available for every possible concept. Even for those that are available, the accuracy can not always be guarenteed to be of a sufficiently high standard [SKO04a]. It is probable that we will be able to automatically annotate video accurately with broad content descriptions in the future due to huge improvements in feature detection performances over the last few years, as outlined by Naphade [NS04] although at present this has still to be realised.

Ontologies already play a vital roles in indexing in the text domain in areas such as medicine (MeSh Medical Subject Headings), biology (gene), and linguistics (WordNet). They enable the processing and sharing of web-based knowledge between applications. An ontology is a set of concepts and their relationships, usually described in the form of a hierarchical tree structure. As a result they

provide a shared and common understanding of a specific domain which can be communicated across various systems or people.

To achieve complete automatic semantic annotation of broadcast TV news, it is necessary to define and standardise an ontology containing a broad set of concepts appropriate for the new broadcast domain as outlined in [Hau04]. An ontology can be constructed either manually, semi-automatically or fully automatically. Semi-automatic construction of ontologies are often restricted in some way to the particular collection they were developed for and as a result tend not to be reusable across different collections. Fully automatic ontology construction requires a large set of concepts, something which is not yet feasible in the broadcast news domain. In the following sections we discuss an ontology that was manually constructed for the broadcast TV news domain. This ontology was used for the annotation of over six thousand keyframes and the resulting manually annotated descriptions provide the input for our concept-based novelty detection model.

RTÉ is Ireland's national television broadcaster and broadcasts three TV channels nationwide. It also has an extensive archive of TV including its own broadcast TV news which goes back several decades. RTÉ manually annotates its own broadcast TV news programmes and other home-produced materials and provides an interesting set of guidelines to annotators for describing the content of a shot [RTE02]. When an annotator is presented with a shot, he must start by annotating the subject of the shot which is followed by a description of the subject's movement (sitting, standing, walking etc). Finally the annotator is required to annotate any secondary subject(s) in the shot. RTÉ annotation guidelines highlight the fact that even though one or more secondary subjects may not be important enough to be retrieved in their own right, the effect of these secondary subjects' presence on the value of the main subject may prove a limiting or enhancing factor during retrieval. RTÉ annotation guidelines suggest that it is important to describe all that is happening in an image so that a person reading the description can visualise the

image and judge whether it is likely to suit their needs.

Current ontologies

TRECVid 2003 instigated a major video annotation collaboration effort resulting in a total of sixty-two hours of video being annotated from the TRECVid 2003 development collection. The ontology used (initially developed for a previous track) consisted of eighty-five semantic labels. Through various additions and deletions over the next two years a final ontology of one hundred and thirty three concepts organised in a hierarchical structure was developed. This ontology consists of thirty-eight scene, thirty-five event, forty-nine object and eleven sound features [LTS03].

During TRECVid 2005, a major collaborative annotation effort was once again accomplished. The Large Scale Concept Ontology for Multimedia (LSCOM) group set out to eventually standardise an ontology of approximately one thousand concepts that will accurately and broadly describe the content in broadcast TV news [Hau04]. LSCOM developed a skeleton ontology from the full 1000-concept LSCOM ontology during development called LSCOMLite, consisting of thirty-nine concepts. It was proposed to divide the semantic space into seven orthogonal dimensions based on Gan's work of "Deciding What's News" [Gan80]. Within each dimension a small number of concepts were assigned with concepts chosen in order to be as broad as possible, while at the same time being possible to detect automatically in video content. They approached the population of ontology construction by adopting a breadth first approach rather than the usual depth first approach.

The ontology developed for the TRECVid video collaboration effort of 2005, LSCOMLite, allows the rapid manual annotation of video and aids search and retrieval. However it is inadequate for other information retrieval tasks including novelty detection as it is very sparsely populated and consequently contains very few discriminating concepts due to a high proportion of the images hav-

ing the same concept sets. Novelty detection requires a set of broad semantic concepts which in effect would allow the user to visualise the image from the annotation alone. LSCOMLite outlines a basis from which an ontology should be developed.

During the construction of an ontology we believe it is necessary to consider the following points namely:

Generality Video is by its nature very diverse in its content, and this can even be seen in broadcast TV news. When designing and creating an ontology to describe this content it is important to recognise this broad domain and so it becomes necessary to choose concepts in such a way that they can be reused over many different queries.

Automatically detectable concepts The purpose of the development of a manual ontology is to standardise the description of video content in order to assist the research and development of specific automatic concept detectors for the defined concepts. It is therefore imperative when designing an ontology and choosing concepts that they are or could possibly be in the future, feasibly detected in video content from the perspective of automatic and semi automatic detection and with a good degree of accuracy.

An even spread across concepts in an ontology An ontology should be designed so that it can be useful for a wide range of applications. As a result, concepts should be chosen in order to avoid the occurrence of Zipf's law distribution over the concepts and increase the amount of good discriminating concepts. Within most text documents discriminating terms occur seldom, while terms containing very little useful or indiscriminate data frequently occur. However within text documents there is little extra cost in having these frequently occurring words since term weighting techniques can easily eliminate their effects. However such a distribution within an ontology would be

very costly as frequently occurring words would offer little discrimination between shots and the annotation of every term within a video collection costs time and effort.

4.3.4 Definition of a New 202-Concept Ontology

It is necessary to provide the correct balance of generality, reusability and broadness/eveness during the creation of any ontology. The ontologies from the LSCOM project meet the requirements of generality, in so far as the ontology concepts are spread across seven dimensions and strive to achieve automatically detectable concepts. However, we believe the ontology is inadequate to support applications such as novelty detection due to the fact that it is too sparsely populated and as a result many shots are annotated with the same high level concepts, even though the visual contents of the images are clearly different. We propose an ontology that builds upon the work of LSCOM. We agree with the division of the semantic space into seven dimensions which is given in table 4.3.4 but feel that the concepts within each dimension should each be expanded using a hierarchical tree structure. This will allow for more detailed content description of a shot.

While the restriction of concept terms to those that are currently obviously feasible to detect is of course a nice idea and a good driver for LSCOM, it is impractical if one needs to create a standard ontology for the future which is movable across other domains. In a few years we can expect feature detectors to have made significant progress in detecting many concepts. As we are interested in working with the resulting data from annotated video, it is not necessarily our priority to satisfy this requirement, although almost all of the concepts within our ontology we assume could be feasibly detected in the future. Our ontology has been developed with both the LSCOM and RTE guidelines in mind.

There are two ways to construct an ontology; the concept-driven approach and the data-driven approach as defined in [JS03]. In the data-driven approach

the ontology is largely constructed from data within the domain, however it is necessary to have some domain knowledge when manually constructing it. The concept-driven approach does not require any data, it is constructed solely on the domain specific or general knowledge of the developers. This approach is more likely to satisfy the requirement that the ontology should be "reusable across different queries" as data from a particular source or data set is not used.

To acquire the set of concepts for our ontology we asked three individuals (all from a non-computing background) to each describe using words, four hours of randomly chosen video from the TRECVid2003 collection represented in shot form. Following the RTE annotation guidelines mentioned earlier and the observations that the most effective concepts are settings and named entities (outlined in [KN04]), annotators were encouraged to firstly describe the subject in the shots and its settings followed by the subjects movements and then any secondary subjects in the shots. This set of words were manually grouped into the seven dimensions defined by LSCOM. Once done, all words were further refined by checking each word against the WordNet lexical database (described in [BMT93]). These words were formed into further clusters using WordNet allowing us to create an hierarchical structure or concept links within the ontology. Consider for example the concepts, car, bus, plane, trucks. All these concepts are a form of "vehicle" and are described as hyponyms of the concept vehicle within WordNet. The resulting ontology, while keeping the seven dimensional space, contains two hundred and two concepts. This can be broken down as shown in Table 4.3.4 into the seven dimensions. All of the concepts which compose the ontology can be seen in their heiracherical structure in the appendices.

4.3.5 Inter-Concept Similarity

Once the ontology has been constructed, it is interesting to evaluate the similarity between concepts within the ontology. Research was carried out within our research group Koskela et al [KSG06], into a model that estimates the "goodness

Dimensions	#concepts LSCOMLite	#concepts
Program category	7	8
Settings/Scene/Site	16	49
People	7	31
Objects	8	52
People Activities	2	41
Events	3	17
Graphs	2	4

Table 4.4: Distribution of Concepts in LSCOMLite and DCU ontology respectively

of a semantic concept model over a clustering in the low level feature space". We evaluated concepts within the LSCOMLite ontology by analysing the way those concepts had been assigned to shots in the collaborative annotation, in an attempt to extract each concept's five most similar concepts in the ontology based on usage. Given each concept in the ontology and its set of five most similar concepts, all randomly ordered, users were asked to manually pick the odd concept from the set. This analysis allows us to identify concepts that are naturally linked to other concepts, while at the same time highlights outliers, concepts that have no obvious similarity or link to any other concepts in the ontology. This analysis helps to give a picture of the overall efficacy or shape of the ontology. This same model was applied to our 202-concept ontology developed for the novelty detection task. The following table, Table 4.3.5, gives an example of some of the concepts, along with their five most similar concepts, chosen randomly from the set of 202 concepts. The full table is available in Appendix B.

Semi-Automatic Annotation

Once an ontology has been developed the next step involves associating various concepts within the ontology to related features in an image or keyframe. Annotating an entire video collection is a very time consuming process and hence it is necessary to have tools which can yield the highest level of quality semantic descriptions within a reasonable period of time. Since its inception in the TRECVid benchmark in 2003 the annotation task has become hugely popular, creating a number of automatic and semi-automatic annotation tools to aid human annotators in annotating a semantic description of video content.

DCU Annotation Tool In order to create a manual annotation and use it for experiments on novelty detection, we developed our own video annotation tool. The DCU annotation tool developed within the Center for Digital Video Processing is an MPEG-7 annotation tool. It takes as input an MPEG-7 video description and corresponding video. It also takes an ontology represented in MPEG-7 format. Whilst quite similar to *VideoAnnEx* [LTS02], this tool has highlighted the fact that better interface design can lead to quicker annotation of a keyframe and hence reduce annotator boredom and frustration. Some notable features of the tool include both an hierarchical tree structure display of the ontology and a "hot" keyword display, allowing a user to quickly navigate though a list of concepts once the first letter is known. The keyframe being annotated is enlarged in the center of the interface allowing the annotator to accurately define what exists in the image. A screengrab from the annotation tool in use can be seen in Figure 4.3. The annotation tool also provides the facility of automatically assigning parent nodes of a specific child concept to an image once selected by the annotator. For example , if an annotator select the concept "car" to identify an object in the image, the DCU annotation tool automatically adds the concepts parent node, "vehicle" to the MPEG7 image description. There is no facility to adapt or customize the ontology within the interface hence avoiding addition of unsupervised concepts to the ontology.

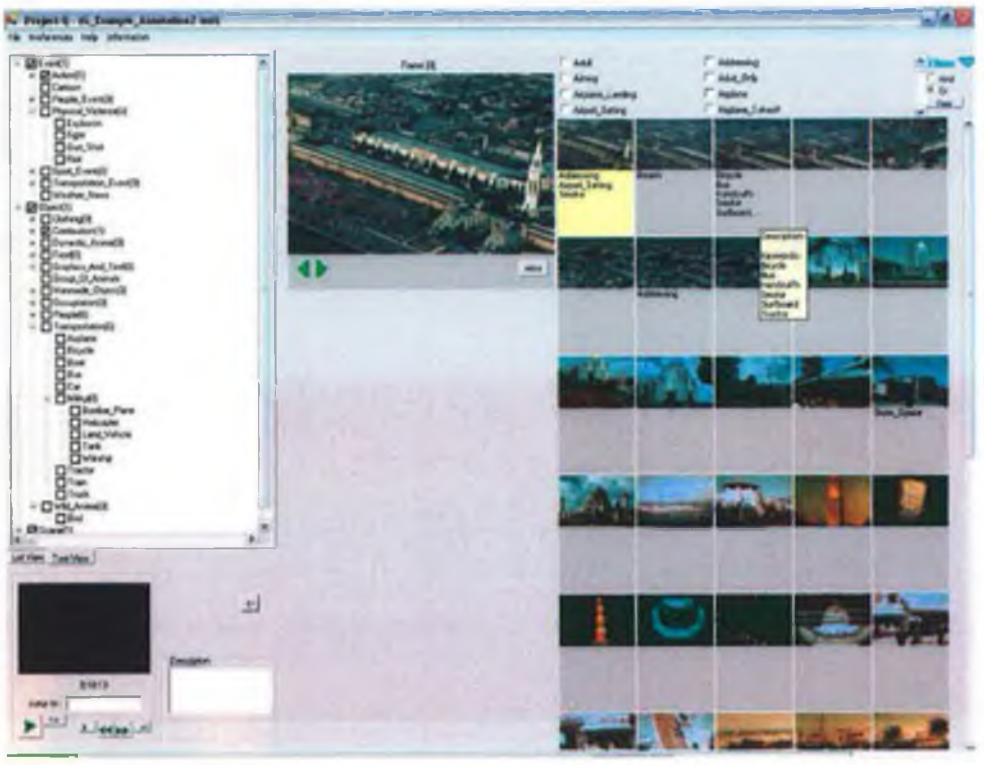

Figure 4.3: The DCU-tool screen dump

This tool was used to annotate the TRECVid2004 collection using our own 202-concept ontology.

4.3.6 Manually annotated novelty detection component

The manual annotations were represented in the form of MPEG-7 descriptions. These descriptions are preprocessed to extract the manually annotated concepts and align each of the concepts to their associated shots. The resulting data is very similar to ASR text portions, in that there are only a few words used to describe a shot. As a result the text novelty component described in section 4.3.1 was applied to the concepts to assess a shot's novelty value when compared to a previously seen shot.

4.3.7 Choosing Threshold Values

As discussed previously the novelty of a document is subjective with different people having different tolerance levels for the existence of redundant data within a results set. Hence novelty threshold values vary from assessor to assessor. In order to control the amount of novel shots to be displayed within a results list we use threshold values which dictate the level of novel data a shot must contain in order to be considered a novel shot. The higher the threshold value, θ, the less tolerant the model is to redundant data. This is particularly suited when we have an information need where we have very little tolerance for sifting through shots containing no new information. Decreasing θ decreases the level of novel data which a shot must contain in order to be considered novel and allows the model to return a greater number of shots as novel. This is more suited to people who don't mind viewing some redundant information in their quest for information. Optimal threshold values θ for novelty detection within video are determined through experimentation. This is further discussed in Chapter 6.

4.3.8 Combining novelty components

It has been seen elsewhere [Hom05] that there are many instances in information retrieval where one modality alone will fail to produce optimal results, however these results can be improved when the correct combination of modalities are used. The final stage of our novelty detection model involves the unification of the various novelty components in order to produce an overall novelty value for each shot. Shots that are above a specified novelty threshold will be highlighted as novel in the list of retrieved shots and highlighted for the user.

The combination of visual feature components is carried out prior to the calculation of an overall novelty score for a shot within a component. This is achieved due the fact that visual features can be normalised within the same ranges. The combinations of visual features include:

- The combination of low-level features. In this case histogram normalisation is first applied to the feature's histogram representations. The similarity measures are then applied for each feature. The similarity values are linearly combined to produce an overall novelty score described in section 4.3.2

- The combination of the automatically detected high-level semantic features, which are combined according to a weighted linear combination, defined by CMU for the various query categories described in section 4.3.3

The process of combining text components with visual components is carried out after the novelty values have been determined for each component independently due to the different domains upon which each is assessed. The unification of text and visual components is accomplished using Boolean logic. A shot is considered novel if, and only if, both the text and visual components agree that the shot is novel, while a shot is considered redundant if either one of the components believes the shot is redundant. These combinations include text combined with low level and text combined with automatic high level concepts.

Finally the combination of the text and the manually annotated semantic concept components is once again carried out prior to the calculation of an overall novelty score for a shot within a component. Combination is achieved by combining the associated ASR transcript portion and associated manually annotated concepts of a shot into an extended text portion. A shot's novelty values is then determined by applying the text novelty detection model.

4.4 Summary

In this Chapter we introduced the idea of novelty detection from within a retrieved results set for any user specific topic in the video domain and more specifically within the broadcast TV news domain. We outlined the need for

novelty detection models when dealing with broadcast news collections. A lot of overlap can occur when the collection contains similar stories, from more than one broadcaster or repetitive video footage due to headlines and summaries repeated within the TV broadcast and across broadcasts. This can lead to a lot of redundant video being presented to the user when he/she requires information on a specific topic.

We discussed various issues which needed to be considered prior to the development of a model for the detection of novel shots from a results list. These considerations included the evolution of news stories, human perception, the overall structure of video and the multiple modalities that can be extracted from a video sequence offering valuable information.

We described the various modalities that are used in the novelty detection model including text; low-level features namely MPEG-7 colour structure and MPEG-7 Edge histograms, HSV colour evidences, Canny edge evidences and Gabor texture evidences; and higher level semantic features when are captured both manually and automatically for each shot in the collection.

We continued by discussing the normalisation of features and the shot to shot similarity distance measures which were chosen for the comparison of shots.

Finally we introduced the novelty detection models designed to accurately identify novel shots from a results list. This model is broken up into four separate novelty components namely, text, low-level features, automatic concepts and manual concept components. Each of these components are capable of determining a shot's novelty value based solely on the evidences contained within the shot and the previously seen shot.

In the next Chapter we discuss the experimental setup for the evaluation of the novelty model developed for the video domain using varying combinations of features.

Concept	Five most similar concept
addressing	roles face sitting indoor politics
adult	person male face people_activities movement
airplane	boat_ship transportation_event sky road water_body
beach	mountain sky cloud commercial water_body
bicycle	road vehicle carrying objects car
bird	animal bicycle interviewing text_overlay greeting
blackboard	hospital_setting doctor emergency_services vehicle school_setting
boat_ship	airplane water_body transportation_event vehicle cloud
bomber_plane	airplane_takeoff transportation_event airplane boat_ship mountain
bottle_drink	female gesturing senior_citizen house_setting telephone
bowing	hospital_setting weapon science_technology doctor gun
bridge	building events road car military_personnel
british_flag	house_setting chair table bottle_drink female
building	news setting_scene_site outdoors standing movement
bus	car_crash emergency_services town_square city_street vehicle
camera	city_street table protesting crowd standing
candle	meeting_board_room talking_speaking bottle_drink newspaper
driver	newspaper store_setting city_street dog greeting
driving	vehicle car road objects carrying
drum	movement city_street standing factory_setting table
eating	house_setting food restaurant_setting bottle_drink senior_citizen
embracing	looking_around objects people_activities factory_setting vehicle
emergency_services	walking_running vehicle car road group
entering	statue_monumoment standing government_leader building presidient
entertainment	male people_activities adult people person
events	sport_event sports playing text_overlay sign
pilot	sky driving road airplane vehicle
playing	sports sport_event tool events sign

Table 4.5: Distribution of Concepts in LSCOMLite and DCU ontology respectively

Chapter 5

Experimental Methodology

In this Chapter we will look at the creation of two video collections and two corresponding novelty ground truth data collections for the task of novelty detection in the video domain. We will discuss the reasons for the development of a video test collection for the novelty detection task given that there are already several video collections widely available within the video retrieval community. In section 5.2 we will describe the generation of the ground truth data used in novelty detection. We will then look at various characteristics of this ground truth data and in section 5.5 we will present our experimental setup for evaluating the performance of our novelty detection approaches introduced earlier in the thesis.

5.1 A Video Test Collection for Novelty Detection

The detection of novel video shots from within a retrieved results set for any user specified topic, is a new research area within the video retrieval community as was stated earlier. The main aim of the task is to accurately and effectively assess the novelty of a shot to the user topic in the context of previously seen shots in the list of shots returned from a retrieval system.

Each year since 2001, the National Institute of Standards and Technology (NIST) in the US make video collections available to participants of the TRECVid benchmarking activity. This enables the common evaluation and cross system comparative analysis of different video retrieval systems [SKO03, SKO04b, Hom05]. This allows participants to evaluate the effectiveness of their systems for efficiently retrieving shots relevant to a particular set of user topics.

The evaluation of the task of novelty detection however cannot be performed on the TRECVid data collections as made available by NIST, as the corresponding relevance judgments provided by NIST treat all relevant shots as equal, no matter what order they are presented to a user and this means that there are significant differences between the tasks of shot retrieval (which TRECVid evaluates) and novelty detection (which it does not). Novelty detection identifies a novel shot given either a previously seen shot (its topic in a sense) or list of shots, while relevance identifies a shot potentially relevant to a user's query. Unlike the detection of relevant shots, where retrieval is performed on an entire video collection, the detection of novel shots is performed on a list of relevant shots returned for a specific topic.

In order to perform effective evaluation of the performances of the proposed novelty detection models for the detection of novel shots from within a list of shots, two tasks are necessary. Firstly we need to create a video test collection that contains a list of relevant shots for each topic, and in our case we compose this collection as a subset of the video used in TRECVid in 2004. The second thing we need to create is the corresponding ground truth data collection, containing shots manually assessed for novelty for each topic in this video collection.

There are several different task within TRECVid carried out annually as discussed in Chapter 2 including the shot boundary detection task, the feature extraction task, and the different kinds of search tasks. The search task is broken up into three different kinds of search approaches. These include the interactive search task, which evaluates the effectiveness of a system to return

relevant shots involving a human interaction within the task; the fully automatic search task, which automatically retrieves shots relevant to a topic given the official topic description; and the manual search task which automatically retrieves relevant shots given a topic, however this task permits the modification of the topic into a suitable representation for the retrieval system. The manual search task simulates a traditional non-interactive retrieval system (e.g. a single iteration with a system like the Google search engine), where the searcher enters his/her information need and is presented with a list of shots relevant to the topic without further interaction with the system.

To accurately simulate a real world situation (while at the same time creating a video collection that allows the accurate evaluation of novelty detection from a retrieved list of relevant shots), our video collection is composed from the results of a search run submission for the manual search task. As described in Chapter 2, TRECVid's manual search task requires each participating group to submit a list of up to 1000 shots they believe relevant for each topic in the collection. These submissions are then manually assessed for relevant shots. We used the results submitted by the best performing group for the manual task in TRECVID2004, specifically one of the IBM Research manual runs, namely the IBM.Manual_ARC run which achieved the highest MAP score of 0.109, to create the video collections for our video shot novelty detection task. The IBM.Manual_ARC run returned shots to a specific topic by using a multi-modal video retrieval system which relied principally on ASR or text retrieval and re-ranked these shots based on a variety of visual features including HSV colour histogram, HSV correlogram, colour moments, colour wavelets, texture, shape and edge.

In the following sub-sections we will describe the topics we have used, and then the creation and attributes of two video sub-collections, henceforth known as video collection Collection_1 and video collection Collection_2. These will enable us to investigate whether our novelty detection models perform consistently across collections. We will also devote a sub-section to describing how our

assessments of novelty were made by our own assessors.

5.1.1 Topics

The TRECVid query topics are used not only by groups participating in the search tasks to formulate queries to their systems but are also used as a reference for the assessors at NIST who determine each shot's relevance to the particular topic in question. Within the novelty detection task, the TRECVid query topics are used as a reference for the assessors to determine whether a shot is novel with regard to a previously seen shot during the creation of the truth data for each topic. The 24 topics, which were provided by NIST as part of the TRECVid2004 collection, simulate a "real world" user information need and are listed in the appendices. Topics can differ in terms of recall where very broad topics can return very many relevant results while narrow topics may have very few video shots that are relevant to the specific information need. Topic 140, for example, "Find shots of one or more bicycles rolling along" contains five relevant shots while Topic 130, "Find shots of a hockey ring with at least one of the nets fully visible from some point" contains 134 relevant shots.

It is more likely that topics which contain very few relevant shots will contain a greater proportion of novel shots. Topics with large numbers of relevant shots are more likely to contain a larger proportion of redundant shots. This will be seen later.

5.1.2 Video Data

The video data used in TRECVid 2004 consisted of broadcast TV news programmes from two different US broadcasters, ABC World News Tonight and CNN Headline News. These news programmes were broadcast over an overlapping a time period, from January to June 1998. This makes it suitable for work on novelty detection since test collections containing any kind of data from an overlapping time period are often more likely to contain redundant or

overlapping data on specific topics. An example of this can be seen in Figure 5.1.2 which displays some of the results (video shots) returned for Topic 133 "Find shots of Saddam Hussein". It was agreed during the manual judgements of novel shots that Shot123_37 from the CNN broadcaster was redundant when Shot120_133 from the ABC broadcaster had been viewed previously.

Figure 5.1: Example of video shot overlap between broadcasters

In addition to raw video footage and keyframes (shown in Figure 5.1.2), the video data collection also contains the Automatic Speech Recognition (ASR) transcripts supplied to all TRECVid participants by LIMSI [JGA02], shot boundary definitions and their representative keyframes, the 24 search topics, a collection of low-level visual features and high level semantic features provided by some of the participants in TRECVid to all other groups. All of these features were originally provided by NIST to participants of the TRECVid2004 conference as part of the video collection. The text transcripts were partitioned with respect to the shot boundaries and each shot was assigned a set of associated (spoken) words. Similarly, low-level feature evidences and high level semantic features were aligned with each shot in the collection.

Video Collection Collection_1: Video collection Collection_1 consists of 579 shots broken up into 375 shots from the ABC news programmes and 204 shots from the CNN news programmes. Table 5.1, displays the collection, partitioned into the twenty four topics. It can be clearly seen that each topic contains a varying number of relevant shots and the reason for this characteristic is described in section 5.1.1.

To ensure as many relevant shots as possible were considered for each topic,

only relevant shots (manually judged as relevant to each specific topic by the NIST assessors) that were submitted by the IBM run for the manual task and subsequently, were considered. These shots were ordered chronologically and added to the collection. Figure 5.2 shows the creation of the Collection_1 video test collection.

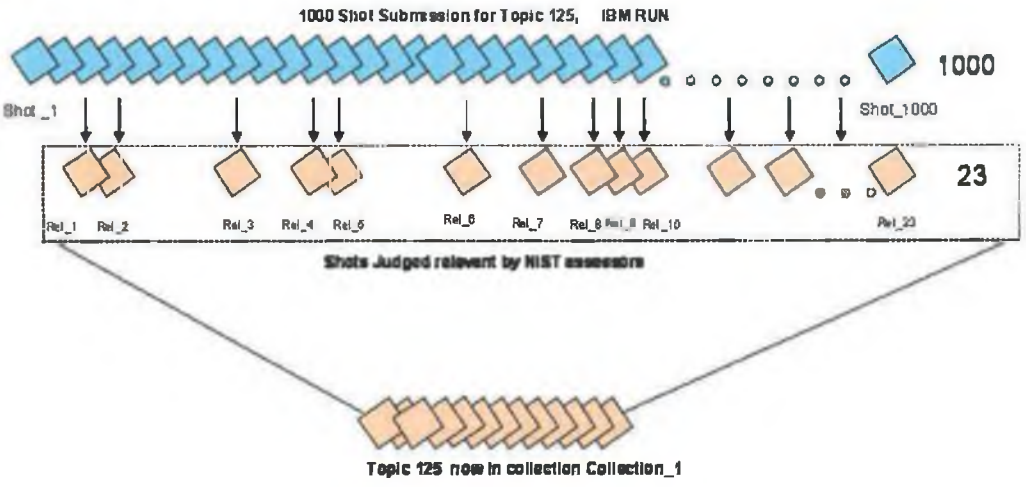

Figure 5.2: Creation of the Collection_1 Video Test Collection

Video Collection Collection_2: Video collection, Collection_2 now consists of 837 shots broken up into 613 shots from the ABC news programmes and 224 shots from the CNN news programmes. Table 5.2, shows how the collection was partitioned into the individual topics and displays the number of shots within each topic. Figure 5.3 shows the creation of the Collection_2 video test collection.

In addition to the data already provided for the TRECVid2004 collection, NIST also provided story boundaries for the ABC and CNN news programmes. Each shot within the original TRECVid2004 was aligned to these story boundaries.

Once a relevant shot has been found within a collection of news programmes for a specific topic it is highly likely that other relevant shots may be found within the same news story. For each shot within each topic in Collection_1, we examined the story it was aligned to. Each story consists of one of more shots

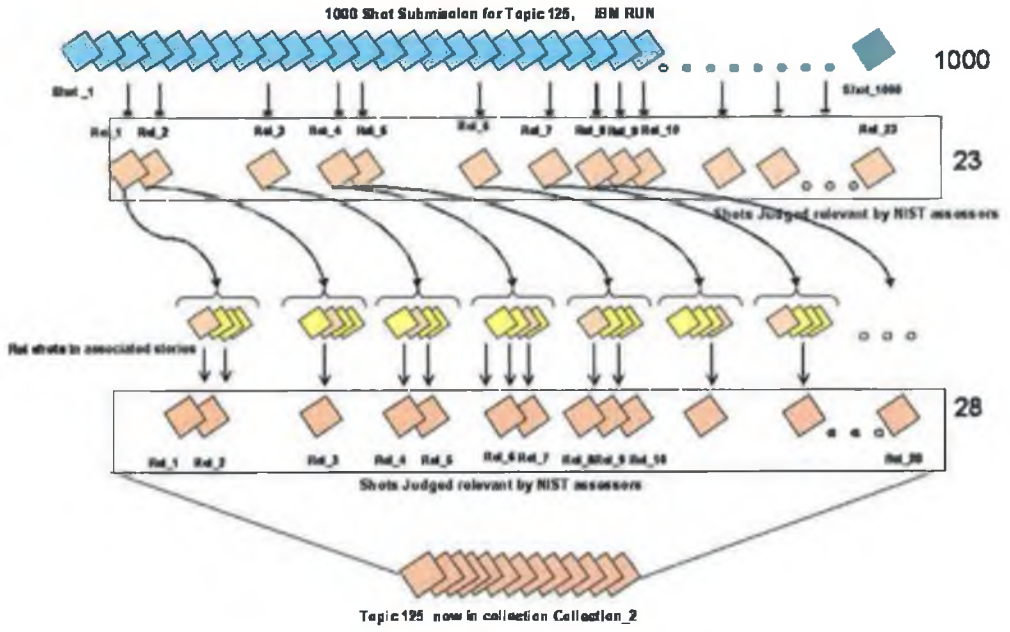

Figure 5.3: Creation of the Collection_2 Video Test Collection

and for each shot in turn we assessed that shot's relevancy to the specific topic using the relevance judgments provided by NIST as part of the TRECVid2004 collection. Each relevant shot found within these stories was added into the Collection_2 collection to the specific topic being investigated.

Collection_2 contains all the shots within the Collection_1 collection, however due to the process described above it contains an additional 258 relevant shots. Therefore Collection_1 is a subset of Collection_2.

As a result each of the two video collections created for the novelty detection task are a subset of the video collection provided to the participants of TRECVid2004 [SKO04b] by NIST. The major difference between the novelty detection video collections, Collection_1 and Collection_2, and a traditional TRECVid video collection, is the partitioning of each collection into twenty four subsets which correspond to the twenty four topics containing relevant shots to each topic, as opposed to a traditional video collection, which is partitioned into a collection of individual news programmes containing shots that

compose the stories for that news programme.

5.2 Novelty Judgments

To obtain accurate experimental results and to allow us to evaluate the performance of the novelty detection approaches presented earlier, it was necessary to create a ground truth data collection or benchmark of unbiased novelty decisions for each video collection that we use here. The creation of the ground truth data for the novelty detection task required that the set of all shots within a results list for each of the 24 topics in each video collection, covering topics 125 to 148, are manually judged to determine whether they are novel or redundant with respect to previously seen shots in the list.

The assessors who performed the novelty detection judgments as part of this work were four postgraduate research students unaffiliated with the current research project. Two assessors performed the novelty detection task on the video collection Collection_1, while two other assessors performed the task on Collection_2. This was done to avoid over-familiarity with the topics and their corresponding sets of relevant shots when determining the novelty of a shot. Each assessor performed the task independently and on a per topic basis.

5.2.1 Assessors Guidelines

The assessor's task was defined as follows; Given a chronologically ordered list of known relevant shots to a particular topic, reduce this list to contain only shots that provide novel information on the topic while at the same time maintaining the original list ordering.

Initially each assessor was given a chronologically ordered list of known relevant shots for each topic as per assumption 1 discussed earlier in Chapter 3. They were instructed to:

- Judge each topic separately.
- Assessors were told that the first shot in the list of shots for a specific topic was always novel. This is in keeping with assumption 6 also discussed in Chapter 3 which states that a user knows nothing about the topic at the time the initial document or shot is displayed and that all information about the topic is gathered as a user progresses through the shot list. Assessors were instructed to always place the first shot into the novel list for the specific topic.
- Assessors were asked to continuously refer back to the original topic definition. This was in an attempt to refresh the actual original information need and help the assessor identify redundant shots already selected which cover that information need.
- Assessors were then instructed to make a decision about a shot's novelty value when compared to shots previously seen from the list up to that point.
- Each novel shot found in the list was added to the novel list for that topic. Each redundant shot found in the list was placed in the redundant list.
- Assessors were instructed to continue the process for each subsequent shot in the relevant list for all topics in the video collection.

Once the novelty judgements had been made for each topic in turn the assessors were not allowed to go back and undo a judgement made earlier. The exact guidelines given to the assessors are available in the Appendix C.

5.2.2 The Assessors

The assessors determined the novelty of a shot based solely on their opinion of what a novel shot should contain for the topic being processed. This models the "real world" situation where each person has their own internal definitions of

novelty and what a shot needs to contain in order for it to be considered novel compared to a previously seen shot. In Chapter 4 section 2.4 we discussed subjectiveness in the perception of visual images [Sha86] and the fact that this can lead to variations between assessors' ground truth data. Different assessors have different opinions on a shot's novelty value, based on what they perceive to be important in the shot. This characteristic was noted during the observation of each pair of assessors during their manual assessment of both video collections. The use of threshold values to determine the level of novelty that should exist in a shot before it is actually considered a novel shot, is a direct attempt at trying to model this human variation in redundancy tolerance (or lack thereof) within models for novelty detection. Yet another observation noted during the manual assessment of novel shots from within a list for a specific topic was the way in which the assessors interacted with the task. The assessors approached the problem by representing the shot under investigation as a query shot in a sense. This query shot was compared against all shots seen previously, on a shot by shot basis, to determine the similarity between them. This is again consistent with human perception of visual images [Sha86, Lay94, Ens95]. It is difficult to make a decision about whether a very similar image has been seen before (and hence to determine a shot's novelty) when comparing a shot against an entire set of shots. Intuitively humans perform the task on a shot to shot basis where a decision can be made directly. If the query shot is similar to a previously seen shot, the assessor makes a decision on its novelty value based on the contents of the shot.

Once the task was completed by each assessor separately, we attempted to eliminate as far as possible disagreements between the assessors. This is a major issue in the creation of ground truth data for novelty detection systems having also been experienced by TREC in the generation of truth data for the novelty track [Har02, SH03, SH04]. Zhang *et al.* [ZCM02] also experienced this problem during the generation of their truth data. Considerable time was put into resolving these assessor judgment differences, however each assessor had their own opinions regarding the novelty of a shot and these differences remained

unresolved. Figure 5.2.2 shows an example of a shot for which both assessors failed to reach agreement on its novelty value given the previously seen set of shots. One assessor believed the shot should be novel as it contained an extra hockey player while the second assessor argued that the shot was redundant as the net had already been shown from that angle in a previous shot.

Figure 5.4: Similar keyframes where assessors disagree over their respective novelty value

In order to overcome this issue and create a truth data as accurately as possible, only the novel shots that have been commonly agreed upon by the **two** assessors for each topic, the intersection of the novelty judgments, are used as the ground truth data in our experiments.

Upon completion of the manual judgement of novel shots within each topic on our two test collections of relevant shots, the shot details were logged to the appropriate topic ground truth data files. Two sets of shots, the novel and redundant sets respectively, now exist for each topic within both video collections.

5.3 Analysis of the Ground Truth

In this section we will look at the characteristics of the truth data, developed for Collection_1 and Collection_2 separately. The assessors who manually created the ground truth data for Collection_1 are henceforth called assessor A and assessor B while the assessors that created the ground truth data for Collection_2 are henceforth referred to as assessor C and assessor D.

Table 5.1 shows the number of relevant shots per topic and the corresponding

number of shots judged as novel by each assessor per topic. Bolded entries in the columns named "Assessor" highlight the assessor (A or B, C or D) who picked the minimum number of novel shots for each topic. As the topics used by us had not been defined or formed by either of the assessors who performed the manual novelty assessment, each assessor should have an independent opinion as to which shots are novel. To achieve this the assessor finding the least number of novel shots per topic was considered the primary assessor for that topic, henceforth called the minimum assessor. The judgements made by the maximum assessor were taken as the human agreement measures for the completion of the novelty detection task. The column named "Total" represents the total number of relevant shots per topic. This value is static over both assessors. The column named "Novel" represents the total number of shots judged novel by the associated assessor for that topic from the total relevant shots. "%Novel" represents the percentage of shots judged novel from the list of relevant shots for that topic. "Intersect" represents the intersection of the novel shots as judged by both assessors. "Overlap", as defined by the TREC novelty track, measures the percentage of matching shots between the set of shots judged as novel by the two assessors, over each topic. "Coverage", also defined by the TREC novelty track, measures the percentage of the minimum assessor's shots that were also chosen by the maximum assessor for a particular topic.

On average there are approximately 25 relevant shots per TRECVid topic of which on average 19 are judged as novel within the ground truth created for the Collection_1 collection. This can be seen in Table 5.1. Similarly there are on average 36 relevant shots per TRECVid topic of which approximately 25 are judged as novel within the ground truth created for the Collection_2 collection, as seen in Table 5.2.

It is obvious from Table 5.1 that there are a large number of topics for which a very high percentage of the relevant shots have been judged as novel by each assessor. The percentage of relevant shots within all topics identified as novel

by the minimum assessor range from 38% to 100% with a median percentage of 81% of relevant shots identified as novel for each topic by the minimum assessor. However there is a lot of variation across the different topics.

This trend is repeated through the truth data for the Collection_2 collection, with the percentage of relevant shots judged novel by the minimum assessor ranging from 22% to 100% and a median percentage over all topics of 80%. Again there is much variation across topics as can be seen in Table 5.2.

If we consider the five different categories to which each topic belongs as discussed earlier in Chapter 4 section 2.0, we can observe in greater detail the characteristics of the truth data by assessing the minimum assessors judgements for each topic. Table 5.3 displays the average number of relevant and novel shots for each topic in each category within the Collection_1 truth data while Table 5.4 displays the average number of relevant and novel shots for each topic within each category within the Collection_2 truth data. This analysis indicates that there is a greater amount of redundancy added to the Collection_2 collection for each of the "People", "Sports", "Other" and "General Object" categories. The "Specific Object" category which consisted of only one topic, namely Topic 129, does not differ between collections as seen in Table 5.1 and Table 5.2

It can be observed from Table 5.1 that assessor A emerges as the minimum assessor over 16 topics while assessor B is the minimum assessor over 8 topics. This suggests that assessor A has less tolerance for the redundant data and hence is stricter on the definition of a shot as novel than assessor B. The different tolerance levels for redundant information within a shot can once again be seen between assessors of the second truth data as depicted in Table 5.2, where assessor D is the minimum assessor for 15 of the 23 topics while assessor C is the minimum assessor for 8 of the 23 topics.

Topic 130 "Find shots of a hockey rink with at least one of the nets fully visible from some point", is a typical example of the differences between assessors'

opinions of novelty. As the Collection_1 collection is a subset of the Collection_2 video collection, we would have expected that shots judged as novel within the truth data of Collection_1 would also be judged as novel within the truth data for Collection_2, along with many other shots as a result of Collection_2's creation process. However this is clearly not the case, as seen when we compare Table 5.1 and Table 5.2 for Topic 130. In fact there are less shots judged novel for Topic 130 in the Collection_2 truth collection than there are in the Collection_1 truth collection. This is a direct result of assessor C's strictness in judging a shot as novel within this topic.

From Table 5.1 we observe that in one topic, namely Topic 141 "Find shots of one or more umbrellas", all relevant shots have been judged as novel by both the minimum and maximum assessors. This holds true in the truth data for Collection_2 as depicted in Table 5.2. In addition, the truth data for Collection_2 also contains two more topics which have all their relevant shots judged as novel by both assessors, namely Topic 142 "Find shots of a tennis player contacting the ball with her or her tennis racket" and Topic 148, "Find shots of one or more signs or banners carried by people at a march or protest". This is an accurate reflection of a real world scenario where it is possible to return information which is all novel in the context of a particular information need to a user.

Analysing the difference of one assessor's opinion of novelty and output for each topic against another assessor's opinion of novelty and its output for each topic, highlights some interesting information. The range in overlap between the two assessors' novel shots varies over all the topics ranging, from 0.61 to 1. This variation is however more obviously seen within the truth data for Collection_2 where the shot overlap ranges from 0.21 to 1. This highlights the difference of opinions between assessors over the novelty value of a shot in a list of relevant shots. The average coverage for both sets of truth data is 0.95. This means that the second assessor has judged 90% of the first assessor's shots correctly. As the novelty detection models were designed to automatically detect the novelty

of a shot in a manner similar to that of a human, this coverage figure is a good indicator of the performance our novelty detection models should achieve to effectively detect novel shots from within a list of relevant shots.

The collection of shots considered novel for each topic by both assessors was defined as the official truth data for each collection.

5.4 Evaluation Metrics

We present the results of various implementations of our novelty detection models by primarily looking at the F-measure which is the official measure used in the TREC novelty track. It focuses on set retrieval, evaluating the quality of the novel set returned. The reader is referred to Chapter 3 where we discussed the F-measure in more detail. As was mentioned in that section, the F-measure does not allow accurate cross system comparisons as the same F-measure score can be achieved using different variations of precision and recall.

As a result we will also present the results of our novelty detection models by looking at the average precision and average recall values over each individual category to which the query topics belong. These measures are analogous to those used in traditional information retrieval evaluation as described in Chapter 1, although novel recall now refers to the proportion of novel shots that are retrieved and precision refers to the proportion of retrieved shots that are novel.

The presentation of the models' optimal precision will be of interest to people wishing to receive as much information about a topic with as little redundant information as possible. However the presentation of the models' optimal F-measure value is more likely to be of interest to people wishing to view a maximum number of novel shots (recall) while at the same time returning the maximum level of precision. It has been observed however that, even though the F-measure is defined as the "harmonic mean between recall and precision", it is correlates closely with recall [SH03]. It has been suggested that the reason

for this characteristic is to do with fact that recall is more consistent across all topics than precision. As a result, to achieve a broad coverage of the model's performance, we will present both precision and F-measure values.

5.5 Systems Evaluation

Our novelty detection models are applied to the list of relevant shots for each topic within the video collection to automatically identify a chronologically ordered list of novel shots that should be returned to the user for each particular topic.

We use a fully automatic experimental setup. We believe that fully automatic experimental runs reduce the noise that can be introduced into an experiment when humans interact with a task. This will provide a more unbiased view of each novelty system's performance. It enables each novelty detection model to be compared more easily against other variations of the novelty detection model and further enables these experiments to be repeated independently within the research community.

The primary objective of the experiments is to identify which combination of resources, or runs, perform the best for the detection of novel shots from within a chronologically ordered relevant list of shots for a topic, across both collections. We investigate the performance of our proposed novelty detection models when they use each of the low-level video features, as a separate run and we also investigate the performance of models when they use a combination of different low-level video features on each of the two video collections, Collection_1 and Collection_2, developed for this novelty detection task. We will, for example, look at how a novelty detection model using a combination of text and a small number of specific automatic high level features, performs against a model using a combination of text and a wide range of manually annotated concepts. These experiments are carried out in an attempt to achieve a wider and hopefully balanced view of the increase or decrease in performance, when using the

different resources and their various combinations in models that accurately detect novel shots from a chronologically ordered relevant list. We will also investigate which threshold values provide the best performance for the detection of novelty when using each of the video resources, when compared against the ground truth data. We will present both the optimised and unbiased threshold values performances for each novelty detection model in each video collection. The unbiased threshold values performances are achieved by firstly finding the optimal thresholds values for the Collection_1 collection and testing them on the Collection_2 collection and vice versa.

As discussed in Chapter 4, each TRECVid topic belongs to one of five individual categories, namely "People", "Specific Object", "General Object", "Sports" and "Other". We will investigate the performances of the different novelty detection models that work best on each of these topic categories. We investigate, for example, whether the novelty detection model using HSV colour or the model using a linear weighted combination of high level features which have been tuned for the "Sports" category will out-perform other models in identifying novel shots within "Sports" topic category.

5.6 Summary

In this Chapter we described the video test collection used in order to perform a novelty detection experiment within the video domain. We outlined the reasons for the development of this video collection. We then proceeded to describe the generation of the test collection. In section 5.3 we performed a detailed analysis of each video collection Collection_1 and Collection_2 including of the development of a the ground truth data set for each collection. We outlined the characteristics of the truth data and highlighted the differences between different assessors' opinions of novelty. We then proceeded to discussed the experimental setup.

In the next Chapter we will describe each of the experimental runs separately

and investigate their performances in detecting novel shots from a list of relevant shots for all topics in the video collection, when compared against the ground truth data.

		Assessor A		Assessor B				
Topic	Total Rel	Novel	%Novel	Novel	%Novel	Intersect	Overlap	Coverage
125	23	19	82.61	**16**	**69.57**	16	0.84	1
126	59	**42**	**71.19**	54	91.53	41	0.75	0.98
127	15	15	100	**12**	**80**	12	0.8	1
128	31	**20**	**64.52**	20	64.52	20	1	1
129	5	4	**80**	5	100	4	0.8	1
130	117	57	48.72	**56**	**47.86**	43	0.61	0.77
131	16	16	100	**15**	**93.75**	15	0.94	1
132	10	**9**	**90**	9	90	9	1	1
133	38	**32**	**84.21**	36	94.74	32	0.89	1
134	14	**12**	**85.71**	12	85.71	12	1	1
135	34	**13**	**38.24**	13	38.24	11	0.73	0.85
136	15	**14**	**93.33**	15	100	14	0.93	1
137	39	**37**	**94.87**	37	94.87	35	0.9	0.95
138	16	**15**	**93.75**	16	100	15	0.94	1
139	13	**12**	**92.31**	12	92.31	12	1	1
140	5	4	**80**	4	80	4	1	1
141	4	4	**100**	4	100	4	1	1
142	9	9	100	**6**	**66.67**	6	0.67	1
143	4	4	100	**3**	**75**	3	0.75	1
144	37	31	83.78	**30**	**81.08**	29	0.91	0.97
145	11	**9**	**81.82**	11	100	9	0.82	1
147	20	**19**	**95**	19	95	19	**1**	1
148	44	**40**	**90.91**	41	93.18	39	0.93	0.98
Total	579	437		446				
Average	25	19	0.76	19	0.76		0.85	0.97

Table 5.1: Analysis of Collection_1 truth data

		Assessor C		Assessor D				
Topic	Total Rel	Novel	%Novel	Novel	%Novel	Intersect	Overlap	Coverage
125	28	24	85.71	**20**	**71.43**	20	0.83	1
126	95	92	96.84	**47**	**49.47**	47	0.51	1
127	28	**16**	**57.14**	18	64.29	16	0.89	1
128	60	41	68.33	**30**	**50**	28	0.65	0.93
129	5	5	100	**4**	**80**	4	0.8	1
130	134	**40**	**29.85**	58	43.28	17	0.21	0.43
131	30	28	93.33	**20**	**66.67**	20	0.71	1
132	13	12	92.31	**11**	**84.62**	11	0.92	1
133	44	37	84.09	**29**	**65.91**	28	0.74	0.97
134	20	17	85	**16**	**80**	15	0.83	0.94
135	45	**10**	**22.22**	13	28.89	10	0.77	1
136	15	15	100	**13**	**86.67**	13	0.87	1
137	62	56	90.32	**52**	**83.87**	52	0.93	1
138	22	**21**	**95.45**	21	95.45	21	1	**1**
139	16	13	81.25	**10**	**62.5**	10	0.77	1
140	10	**8**	**80**	8	80	8	1	1
141	4	**4**	**100**	4	100	4	1	1
142	12	**12**	**100**	12	100	12	1	1
143	11	10	90.91	**9**	**81.82**	8	0.73	0.89
144	48	38	79.17	**35**	**72.92**	32	0.78	0.91
145	25	21	84	**20**	**80**	18	0.78	0.9
147	23	23	100	**20**	**86.96**	20	0.87	**1**
148	87	**87**	**100**	87	100	87	1	1
Total	837	630		557				
Average	36	27	.75	24	.66		0.8	0.95

Table 5.2: Analysis of Collection_2 truth data

Topic Category	Avg. Total Relevant Shots	Avg. Novel Relevant Shots	%Novel
People	31	21	70
Specific Obj.	5	4	80
General Obj.	7	6	86
Sports	47	25	53
Scene (other)	28	23	82

Table 5.3: Analysis of Topic Categories within the Collection_1 truth data

Topic Category	Avg. Total Relevant Shots	Avg. Novel Relevant Shots	%Novel
People	43	24	58
Specific Obj.	5	4	80
General Obj.	11	8	72
Sports	54	22	41
Scene (other)	45	33	73

Table 5.4: Analysis of Topic Categories within the Collection_2 truth data

Chapter 6

Experimental Results

In this chapter we report the results of the experiments carried out using each of the individual approaches for the detection of novel shots in a results list on each of the two video collections, Collection_1 and Collection_2, developed for the novelty detection task in the video domain. The experimental results are compared against the baseline novelty performance, a system returning all relevant shots as being novel to each topic.

6.1 Experimental Results

In this thesis we have introduced the concept of novelty detection from a chronologically ordered list of shots known to be relevant for a particular topic. Through experimentation and analysis of the results in this chapter, we aim to answer the research questions posed at the beginning of the thesis namely;

1. Can novel shots be automatically detected from within a list of shots within the video domain ?

2. Do models designed to detect novel shots from a chronologically ordered list of shots using text resources alone out-perform other resources and

combinations of resources also available within the video domain or does novelty detection need to utilise the other resources available from within video to accurately complete the task ?

3. How do novelty detection models developed for the identification of novel shots from a chronologically ordered list of relevant shots for a topic within the video domain, perform compared to a human assessor's performance of the task ?

4. How do the performances of the many modalities available for each video sequence compare to each other in the task of detecting novel shots from a chronologically ordered list of relevant shots for a topic ?

A series of controlled experiments were carried out on each of the four resources attributed to a video, namely, text taken from automatic speech recognition (ASR), low-level features (such as colour and texture), high level semantic features and manually annotated concepts on both test collections, Collection_1 and Collection_2 which were described in Chapter 5 section 1, under the same experimental conditions. This allows us to accurately explore the effect of using different types of video resources in detecting novel shots from a results list and also to explore the effect of using combinations of these resources to find novel shots.

These experiments enable us to accurately compare the performances of each of our novelty model approaches against a human assessor's performance and also against a baseline that returns every shot in a results list as novel for the novelty task. We analysed the models performance on 23 topics (Topic 146 had no relevant shots) using the ground truth which was developed for each test collection, as described earlier in Chapter 5 section 2.

6.2 Presentation of Results

In this chapter we investigate the performances of each resource, which corresponds to each of the four novelty detection models described in Chapter 4, in accurately identifying novel shots from a results list. These models include

1. Video Novelty Model using text in the form of ASR.
2. Video Novelty Model using low level features.
3. Video Novelty Model using automatic high level concepts.
4. Video Novelty Model using manual annotated concepts.

Each of the four main video resources can be further broken into individual evidences, for example colour and edges in the low level feature category. Each of these evidences offer different information that can affect the performance of the novelty detection models on each of the topics. Consequently each of the four sections are subdivided into five topic category subsections and we investigate the performances of each of the individual feature evidences on these topics. Henceforth we will refer to the performance of a model as the performance of a run.

Within each category subsection, we present the three best performing or optimal F-measure values (Fscores) and their corresponding precision, recall and threshold values of each run for each category within each collection. In addition we present the unbiased results. The unbiased results for Collection_1 are acquired by extracting the Fscores and their corresponding precision and recall figures for the three threshold values that produced the optimal Fscores on Collection_2 and vice-versa. We compare the performance of each run within each collection to the baseline performances for that collection. Table 6.1 and Table 6.2 display the baseline and assessor performances for each topic category over Collection_1 and Collection_2 respectively. When we refer to the baseline figures within the chapter we are referring to each of these Tables respectively.

A similar analysis was carried out for the optimal precision values. However in this case, the optimal precision values have the lowest recall and Fscore and as a results we do not discuss them in this chapter

An analysis subsection will follow each of the four resource subsections identifying the best performing run of that resource for each topic category.

Finally the chapter will contain an overall analysis section which will outline the best performing resource(s), which work well over each of the topic categories. It will contain a subsection which will display the median difference graphs of the best performing runs over each of the topic categories for each of the video resources. Median difference graphs show the per-topic difference between the optimal Fscore achieved by the run under investigation and the median Fscore for that topic. These graphs are used to visually present the performance of the individual runs and allow us to see what types of topics the run can handle well. They will allow us to identify which models are superior for the different topics. The difference between the optimal Fscore for a run and the median Fscore is calculated as

$$Diff = Fscore_{optimal} - Fscore_{median} \tag{6.1}$$

A positive $Diff$ value indicates that the run's optimal Fscore is performing better than the median and conversely if $Diff$ is negative the model is performing below the median. This will enable us to further analyse how each of the runs perform over each of the individual topics.

6.2.1 Topic Categories

Topics vary in terms of both information need and the number of novel shots associated with each topic in the results lists. Averaging the evaluation measures over all topics disguises how each novelty detection model performs on each of the different types of topics and smoothes over any abnormalities in topic performances. This hampers accurate comparison of the performance of models over different topics.

	Prec	Recall	Fscore
All Topics			
Baseline	0.79	1	0.872
Assessor A	0.93	1	0.962
Assessor B	0.94	1	0.967
General Object			
Baseline	0.865	1.000	0.926
Assessor A	0.958	1.000	0.976
Assessor B	0.970	1.000	0.983
Other			
Baseline	0.844	1.000	0.911
Assessor A	0.933	1.000	0.964
Assessor B	0.950	1.000	0.972
People			
Baseline	0.725	1.000	0.823
Assessor A	0.957	1.000	0.976
Assessor B	0.943	1.000	0.969
Specific Object			
Baseline	0.800	1.000	0.889
Assessor A	1.000	1.000	1.000
Assessor B	0.800	1.000	0.889
Sports			
Baseline	0.657	1.000	0.768
Assessor A	0.807	1.000	0.887
Assessor B	0.900	1.000	0.945

Table 6.1: Baseline performances over all categories over Collection_1

	Prec	Recall	Fscore
All Topics			
Baseline	0.71	1	0.808
Assessor C	0.85	1	0.908
Assessor D	0.93	1	0.957
General Object			
Baseline	0.787	1.000	0.876
Assessor C	0.892	1.000	0.940
Assessor D	0.965	1.000	0.981
Other			
Baseline	0.751	1.000	0.847
Assessor C	0.846	1.000	0.907
Assessor D	0.984	1.000	0.992
People			
Baseline	0.598	1.000	0.725
Assessor C	0.848	1.000	0.915
Assessor D	0.920	1.000	0.957
Specific Object			
Baseline	0.800	1.000	0.889
Assessor C	0.800	1.000	0.889
Assessor D	1.000	1.000	1.000
Sports			
Baseline	0.667	1.000	0.718
Assessor C	0.763	0.980	0.837
Assessor D	0.760	0.980	0.811

Table 6.2: Baseline performances over all categories over Collection_2

Topics can be categorised into five different classes such as General Object, Specific Object, People, Other and Sports. Each of the topics which are used in our experiments belong to only one of the predefined categories described in Chapter 4 section 2.5.

As a result, in the presentation and analysis of the results of our novelty detection we look at the results of each run over all topics but in addition we will also look at the topics in each of their respective categories and analyse the performance of each run on each of the categories separately. Dividing topics into categories enables our investigation into which of the different novelty detection models work best on each of these topic categories.

In Chapter 5 during the analysis of the ground truth we observed that some topics contained very little novel shots or in other words contained a lot of redundant shots, while others contained a very high percentage of novel shots. In test Collection_1 the truth data identifies one topic, namely Topic 141 of the 23 topics where all shots within the results set were considered novel. In test Collection_2 the truth data identifies three topics, namely Topic 141, Topic 142 and Topic 148 of the 23 topics where all shots were considered novel. This is the nature of novelty detection in synthetic test collections and has been observed in the TREC novelty track [Har02] and by Allan *et. al* [AWB03]. These topics offer little, in evaluating a model's performance for removing redundant information from a results list because all shots are novel.

Topics where 50% or less of the shots are considered novel in the results set, are of particular interest to novelty detection models, as we can analyse the performance of the novelty model in handling novelty detection over these topics. Table 6.3 and Table 6.4 displays the topics in terms of the percentages of shots that were identified as novel, in the results list for ground truth of Collection_1 and Collection_2, respectively.

< 50%	50-70%	70-90%	90-100%
Topic 130	Topic 125	Topic 127	Topic 131
Topic 135	Topic 126	Topic 129	Topic 132
	Topic 128	Topic 133	Topic 136
	Topic 142	Topic 134	Topic 138
		Topic 137	Topic 139
		Topic 140	Topic 141
		Topic 143	Topic 147
		Topic 144	
		Topic 145	
		Topic 148	

Table 6.3: Percentages of shots found novel in each topic in Collection_1

< 50%	50-70%	70-90%	90-100%
Topic 126	Topic 127	Topic 125	Topic 138
Topic 128	Topic 131	Topic 129	Topic 141
Topic 130	Topic 133	Topic 132	Topic 142
Topic 135	Topic 139	Topic 134	Topic 148
	Topic 144	Topic 136	
		Topic 137	
		Topic 140	
		Topic 143	
		Topic 145	
		Topic 147	

Table 6.4: Percentages of shots found novel in each topic in Collection_2

6.3 Video Novelty Model using Text

During the TRECVid search task the unit of retrieval for the topic is the shot. Over the years retrieval based on text has proven to be the primary initial method of retrieval for those approaches that work best. Systems failing to utilise this important resource usually produce poor performance results in comparison to those systems that do [SO02, SKO03, SKO04b, SKO05]. As a result and also due to the fact that novelty detection was first introduced within the text domain, we investigate a novelty detection model designed to utilise ASR from within video proposed in Chapter 4. The run *ASR_Shot_by_Shot* was used to investigate "the shot by shot" approach to novelty detection when using ASR transcripts for a shot. Run *ASR* was used to explore the performance of the novelty model utilising ASR transcripts for a shot and keeping an accumulative history of all shots seen so far. The question now follows, how well does utilisation of text resources from within video perform in the identification of novel shots from within a chronological list of relevant shots ?

Tables 6.5 and 6.6 display the three optimal and unbiased Fscores achieved by each of the ASR runs over all topics in both Collection_1 and Collection_2 respectively.

	Optimised Results				**Unbiased Results**			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR	0	0.81	0.98	0.872	0	0.81	0.98	0.872
	0.5	0.81	0.98	0.87	0.5	0.81	0.98	0.87
	0.7	0.81	0.98	0.869	0.1	0.81	0.98	0.872
ASR_Shot_by_Shot	0	0.83	0.84	0.819	0	0.83	0.84	0.819
	0.1	0.81	0.73	0.756	0.1	0.81	0.73	0.756
	0.2	0.84	0.59	0.674	0.2	0.84	0.59	0.674

Table 6.5: Results of the Novelty detection model using ASR over *all topics* in Collection_1

Firstly if we consider how ASR performs over all topics in Collection_1, Table 6.5, we observe that "ASR" is performing similar to the baseline which means

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR	0	0.71	0.98	0.8	0	0.71	0.98	0.8
	0.1	0.72	0.96	0.796	0.7	0.72	0.95	0.792
	0.5	0.72	0.95	0.793	0.5	0.72	0.95	0.793
ASR_Shot_by_Shot	0	0.73	0.81	0.75	0	0.73	0.81	0.75
	0.1	0.71	0.75	0.716	0.1	0.71	0.75	0.716
	0.2	0.72	0.65	0.66	0.2	0.72	0.65	0.66

Table 6.6: Results of the Novelty detection model using ASR over *all topics* in Collection_2

that the run is returning all the shots as novel for each topic, or in other words is having no effect on the detection of novel shots. From Table 6.6 we observe that the same run is performing below the baseline novelty performance figures over all topics in Collection_2 suggesting that ASR is actually having a negative effect on the detection of novel shots from within a collection of visual shots.

6.3.1 "General Object" Topic Category

Tables 6.7 and 6.8 display both the optimal and unbiased F-measure values of the novelty run using ASR over all topics in the "General object" category from Collection_1 and Collection_2 respectively.

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR	0	0.898	0.945	0.913	0.0	0.898	0.945	0.913
	1.3	0.897	0.930	0.905	1.4	0.897	0.930	0.905
	1.9	0.897	0.888	0.874	0.1	0.898	0.945	0.913
ASR_Shot_by_Shot	0	0.912	0.903	0.894	0.0	0.912	0.903	0.894
	0.1	0.888	0.810	0.833	0.1	0.888	0.810	0.833
	0.2	0.920	0.658	0.759	0.2	0.920	0.658	0.759

Table 6.7: Results of the Novelty detection model using ASR over the "General Object" topic category within Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR	0	0.785	0.990	0.871	0.000	0.785	0.990	0.871
	1.4	0.822	0.932	0.863	1.300	0.815	0.932	0.858
	0.1	0.815	0.932	0.858	1.900	0.822	0.890	0.831
ASR_Shot_by_Shot	0	0.820	0.890	0.844	0.000	0.820	0.890	0.844
	0.1	0.800	0.823	0.800	0.100	0.800	0.823	0.800
	0.2	0.793	0.757	0.754	0.200	0.793	0.757	0.754

Table 6.8: Results of the Novelty detection model using ASR over "General Object" topic category within Collection_2

From Table 6.7 we can see both runs, "ASR" and "ASR_Shot_by_Shot" perform below the baseline performance over Collection_1 and this is consistent over Collection_2. Table 6.8 suggesting that ASR does not aid in the detection of novel shots in the "General object" topic categories.

6.3.2 "Other" Topic Category

Tables 6.9 and 6.10 display both the optimal and unbiased F-measure values of the novelty run using ASR over all topics in the "Other" category of topics from Collection_1 and Collection_2 respectively.

General Object

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR	0.9	0.854	0.989	0.915	0.1	0.853	0.993	0.915
	0	0.853	0.993	0.915	0.8	0.853	0.993	0.915
	1.1	0.860	0.969	0.909	0.0	0.853	0.993	0.915
ASR_Shot_by_Shot	0	0.859	0.803	0.830	0.0	0.859	0.803	0.830
	0.1	0.863	0.706	0.771	0.1	0.863	0.706	0.771
	0.2	0.853	0.573	0.672	0.2	0.853	0.573	0.672

Table 6.9: Results of the Novelty detection model using ASR over "Other" topic category within Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR	0.1	0.753	0.966	0.836	0.0	0.744	0.973	0.834
	0.8	0.753	0.963	0.835	0.9	0.753	0.963	0.835
	0	0.744	0.973	0.834	1.1	0.756	0.946	0.831
ASR_Shot_by_Shot	0	0.777	0.830	0.798	0.0	0.777	0.830	0.798
	0.1	0.757	0.757	0.750	0.1	0.757	0.757	0.750
	0.2	0.754	0.674	0.691	0.2	0.754	0.674	0.691

Table 6.10: Results of the Novelty detection model using ASR over "Other" topic category within Collection_2

Utilizing ASR within novelty models for the detection of novel shots in the "Other" category over Collection_1 achieves an Fscore of 0.915, an improvement of 0.4% on the baseline, while the corresponding precision value achieves an improvement of 1.2% on the baseline precision value, Table 6.9. The run that accesses novelty based on the shot by shot comparisons performs below the

baseline. If we look at Table 6.10 we see that both runs perform below the baseline in collection_2. This inconsistency is a direct result of the additional shots in Collection_2.

6.3.3 "People" Topic Category

Tables 6.11 and 6.12 display both the optimal and unbiased F-measure values of the novelty run using ASR over all topics in the "People" category from Collection_1 and Collection_2 respectively.

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR	0	0.738	0.995	0.828	0.0	0.738	0.995	0.828
	0.5	0.738	0.985	0.824	0.1	0.738	0.995	0.828
	0.6	0.737	0.980	0.821	0.5	0.738	0.985	0.824
ASR_Shot_by_Shot	0	0.773	0.783	0.760	0.0	0.773	0.783	0.760
	0.1	0.738	0.627	0.672	0.2	0.787	0.520	0.606
	0.2	0.787	0.520	0.606	0.1	0.738	0.627	0.672

Table 6.11: Results of the Novelty detection model using ASR over "People" topic category within Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR	0	0.598	0.997	0.723	0.0	0.598	0.997	0.723
	0.1	0.602	0.960	0.717	0.5	0.597	0.938	0.706
	0.5	0.597	0.938	0.706	0.6	0.597	0.935	0.705
ASR_Shot_by_Shot	0	0.597	0.705	0.630	0.0	0.597	0.705	0.630
	0.2	0.658	0.598	0.619	0.1	0.597	0.668	0.618
	0.1	0.597	0.668	0.618	0.2	0.658	0.598	0.619

Table 6.12: Results of the Novelty detection model using ASR over "People" topic category within Collection_2

If we look at Table 6.11 we see ASR achieved an Fscore of 0.828, an increase of 0.6% on the baseline while the corresponding precision value of 0.738 gives an

increase of 1.8% on the baseline. Table 6.12 however shows that ASR performs badly on the people category over Collection_2 achieving an Fscore of 0.723, a decrease of 0.3% on the baseline performance.

6.3.4 "Specific Object" Topic Category

Tables 6.13 and 6.14 display both the optimal and unbiased F-measure values of the novelty run using ASR over all topics in the "Specific Object" topic category from Collection_1 and Collection_2 respectively.

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	11	1.000	0.750	0.857	11.0	1.000	0.750	0.857
	4.9	0.750	0.750	0.750	4.9	0.750	0.750	0.750
ASR_Shot_by_Shot	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.1	0.750	0.750	0.750	0.1	0.750	0.750	0.750
	0.3	1.000	0.500	0.667	0.3	1.000	0.500	0.667

Table 6.13: Results of the Novelty detection model using ASR over "Specific Object" topic category within Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	11	1.000	0.750	0.857	11.0	1.000	0.750	0.857
	4.9	0.750	0.750	0.750	4.9	0.750	0.750	0.750
ASR_Shot_by_Shot	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.1	0.750	0.750	0.750	0.1	0.750	0.750	0.750
	0.3	1.000	0.500	0.667	0.3	1.000	0.500	0.667

Table 6.14: Results of the Novelty detection model using ASR over "Specific Object" category within Collection_2

From Table 6.13, we see that both runs are achieving a novelty performance similar to the baseline over Collection_1. This means that essentially all documents

are being considered novel by ASR. This is also the case over the specific category within Collection_2 where ASR is again performing similar to the baseline, Table 6.14.

6.3.5 "Sports" Topic Category

Tables 6.15 and 6.16 display both the optimal and unbiased F-measure values of the novelty run using ASR over all topics in the "Sports" topic category from Collection_1 and Collection_2 respectively.

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR	1.7	0.663	0.983	0.773	0.7	0.660	0.993	0.769
	1.6	0.663	0.983	0.772	0.1	0.657	0.993	0.768
	2.3	0.663	0.970	0.772	1.0	0.660	0.983	0.768
ASR_Shot_by_Shot	0	0.697	0.830	0.741	0.1	0.697	0.830	0.741
	0.2	0.810	0.630	0.682	0.0	0.697	0.830	0.741
	0.3	0.807	0.487	0.563	0.2	0.810	0.630	0.682

Table 6.15: Results of the Novelty detection model using ASR over "Sports" topic category within Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR	0.7	0.667	0.973	0.708	1.7	0.663	0.913	0.697
	0.1	0.667	0.973	0.707	1.6	0.663	0.913	0.696
	1	0.667	0.953	0.705	2.3	0.663	0.913	0.700
ASR_Shot_by_Shot	0.1	0.653	0.740	0.654	0.0	0.647	0.720	0.646
	0	0.647	0.720	0.646	0.2	0.670	0.523	0.513
	0.2	0.670	0.523	0.513	0.3	0.643	0.420	0.414

Table 6.16: Results of the Novelty detection model using ASR over "Sports" topic category within Collection_2

From Table 6.15 we can see that ASR achieves an Fscore of 0.773, an increase of 0.7 % on the baseline Fscore while its corresponding precision value of 0.663 gives an improvement of 0.9 % over the baseline over the sports category. However this is not consistent over the sports category within Collection_2 with ASR performing below the baseline performances as clearly seen in Table 6.16.

6.3.6 Summary analysis for text features

As ASR is the primary resource used in shot retrieval it would be expected to perform well during the detection of novel shots from within a list of shots, however as the findings presented above illustrate, this is not actually the case.

We observe inconsistencies between the two collections and these are a direct result of the additional ASR portions associated with the additional shots which are contained within Collection_2. As described in Chapter 5, Collection_2 was firstly composed of all relevant shots for each topic as in Collection_1 however, in addition to these shots, Collection_2 contained relevant shots from each of the stories associated with each of the original relevant shots. This resulted in many shots from Collection_2 containing ASR portions which are connected to the same story. As a result ASR portions may be very similar. Also, we must note that if a shot did not contain an ASR portion, it was considered novel by default.

The accuracy of determining the novelty of a shot using ASR is inconsistent over all topics and in many cases returns all shots as novel or performs worse than the baseline. As a result ASR should not be considered solely in determining the novelty value of a shot within a topic.

6.4 Video Novelty Model using Low Level Features:

In this section we will present the approaches to novelty detection, that utilise low-level features taken from video shot keyframes. Ten different low level features and seven combination variations of these features were investigated to determine the benefit of each feature in assessing the novelty value of a shot. Each variation is represented by a self explanatory run name, which will be used for the identification of the particular approach for the duration of the thesis. We look at two colour features including HSV and MPEG7 colour structure, two edge features including Canny edge and the MPEG7 edge histogram and finally we look at Gabor texture.

We investigate the performance of our proposed novelty detection models when they use each of the low-level video features as a separate run, and we also investigate the performance of models when they use a combination of different low-level video features on each of the two video collections, Collection_1 and Collection_2.

We investigate what effect combining the low-level feature with text (ASR) has on novelty detection performances over all topics and over each of the topic categories separately. The low-level features we use in the experiments have been applied to keyframes extracted from each of the video shots in the collections.

From Tables 6.17 and 6.18 we can clearly see that the two highest performing novelty runs over all topics in Collection_1 include, colour structure "ColourStruc" achieving an Fscore of 0.893 with a corresponding precision value of 0.86, an improvement of 2.4% and 8.7% on the baseline figures respectively and the combination of colour structure and edge histograms "ColourStruc_EdgeHist" which achieves a similar Fscore of 0.893, however the precision is slightly less at 0.84, an improvement of 6.3% on the baseline precision figure. We observe that both low level edge features, edge histograms and Canny edge also perform slightly above the baseline respectively. We also observe that ASR reduces the performance of the highest performing runs within this collection while having no effect on each of the other low level runs over all topics. From Tables 6.19 and 6.20 we can clearly see that this trend is consistent over all topics in Collection_2, with colour structure "ColourStruc" and the combination of colour structure and edge histograms "ColourStruc_EdgeHist" both achieving the highest novelty Fscore of 0.822 an improvement of 1.7% on the baseline figures and precision of 0.74. We note once again that ASR reduces the novelty performance on all runs over all topics in Collection_2.

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
EdgeHist	0.3	0.81	0.99	0.878	0.3	0.81	0.99	0.878
	0.2	0.79	1	0.874	0.1	0.79	1	0.872
	0	0.79	1	0.872	0.4	0.86	0.89	0.865
HSVColour	0	0.78	0.93	0.813	0	0.78	0.93	0.813
	0.1	0.94	0.22	0.339	0.1	0.94	0.22	0.339
	0.2	1	0.17	0.284	0.2	1	0.17	0.284
HSVColour_CannyEd	0.2	0.79	0.92	0.819	0.2	0.79	0.92	0.819
	0.1	0.79	0.93	0.817	0.1	0.79	0.93	0.817
	0	0.78	0.93	0.813	0	0.78	0.93	0.813
HSVColour_CannyEd_ Texture	0	0.79	0.95	0.837	0	0.79	0.95	0.837
	0.1	0.79	0.95	0.835	0.1	0.79	0.95	0.835
	0.3	0.8	0.89	0.813	0.3	0.8	0.89	0.813
HSVColour_Texture	0	0.79	0.94	0.824	0	0.79	0.94	0.824
	0.1	0.9	0.47	0.577	0.1	0.9	0.47	0.577
	0.2	0.94	0.35	0.48	0.2	0.94	0.35	0.48
CannyEd	0.2	0.81	0.99	0.884	0.2	0.81	0.99	0.884
	0.1	0.8	1	0.877	0.1	0.8	1	0.877
	0	0.79	1	0.872	0	0.79	1	0.872
CannyEd_Texture	0.2	0.8	1	0.876	0.2	0.8	1	0.876
	0.1	0.79	1	0.873	0.1	0.79	1	0.873
	0	0.79	1	0.872	0.3	0.8	0.95	0.861
Texture	0	0.79	1	0.872	0	0.79	1	0.872
	0.1	0.89	0.44	0.558	0.1	0.89	0.44	0.558
	0.2	0.92	0.31	0.439	0.2	0.92	0.31	0.439
ColourStruc	0.3	0.86	0.94	0.893	0.3	0.86	0.94	0.893
	0.2	0.81	0.99	0.883	0.2	0.81	0.99	0.883
	0.1	0.8	1	0.877	0	0.79	1	0.872
ColourStruc_EdgeHist	0.7	0.84	0.97	0.893	0.5	0.8	1	0.88
	0.8	0.87	0.93	0.891	0.4	0.8	1	0.878
	0.6	0.82	0.99	0.884	0.6	0.82	0.99	0.884

Table 6.17: Results of the Novelty detection model using low level features for *all topics* over Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_HSVColour_ CannyEd	0.0 & 0.2	0.79	0.92	0.819	0.0 & 0.2	0.79	0.92	0.819
	0.2 & 0.2	0.81	0.9	0.817	0.0 & 0.0	0.78	0.93	0.813
	0.6 & 0.2	0.81	0.9	0.815	0.2 & 0.0	0.8	0.91	0.813
ASR_HSVColour	0.2 & 0.0	0.8	0.91	0.813	0.0 & 0.0	0.78	0.93	0.813
	0.6 & 0.0	0.8	0.91	0.811	0.6 & 0.0	0.8	0.91	0.811
	0.8 & 0.0	0.8	0.9	0.809	0.4 & 0.0	0.8	0.91	0.813
ASR_HSVColour_ Texture	0.2 & 0.0	0.81	0.92	0.824	0.0 & 0.0	0.79	0.94	0.824
	0.6 & 0.0	0.81	0.91	0.822	0.4 & 0.0	0.81	0.92	0.824
	0.8 & 0.0	0.8	0.91	0.82	0.6 & 0.0	0.81	0.91	0.822
ASR_ColourStruc_ Texture_CannyEd	0.2 & 0.0	0.81	0.93	0.837	0.0 & 0.0	0.79	0.95	0.837
	0.0 & 0.2	0.8	0.94	0.835	0.2 & 0.0	0.81	0.93	0.837
	0.2 & 0.2	0.81	0.92	0.833	0.0 & 0.2	0.8	0.94	0.835
ASR_CannyEd	0.0 & 0.2	0.81	0.99	0.884	0.0 & 0.2	0.81	0.99	0.884
	0.2 & 0.2	0.83	0.98	0.881	0.0 & 0.0	0.79	1	0.872
	0.6 & 0.2	0.83	0.97	0.879	0.6 & 0.2	0.83	0.97	0.879
ASR_Texture	0.2 & 0.0	0.81	0.98	0.872	0.2 & 0.0	0.81	0.98	0.872
	0.6 & 0.0	0.81	0.98	0.87	0.6 & 0.0	0.81	0.98	0.87
	0.8 & 0.0	0.81	0.97	0.868	0.0 & 0.0	0.79	1	0.872
ASR_Texture_ CannyEd	0.0 & 0.2	0.8	1	0.876	0.0 & 0.2	0.8	1	0.876
	0.2 & 0.2	0.82	0.98	0.875	0.2 & 0.2	0.82	0.98	0.875
	0.6 & 0.2	0.82	0.97	0.873	0.0 & 0.0	0.79	1	0.872
ASR_ColourStruc	0.0 & 0.2	0.81	0.99	0.883	0.0 & 0.2	0.81	0.99	0.883
	0.2 & 0.2	0.82	0.97	0.88	0.2 & 0.2	0.82	0.97	0.88
	0.6 & 0.2	0.82	0.97	0.878	0.0 & 0.0	0.79	1	0.872
ASR_ColourStruc_ EdgeHist	0.0 & 0.8	0.87	0.93	0.891	0.0 & 0.6	0.82	0.99	0.884
	0.0 & 0.6	0.82	0.99	0.884	0.0 & 0.4	0.8	1	0.878
	0.2 & 0.8	0.87	0.91	0.881	0.0 & 0.2	0.79	1	0.872
ASR_EdgeHist	0.0 & 0.2	0.79	1	0.874	0.0 & 0.2	0.79	1	0.874
	0.0 & 0.0	0.79	1	0.872	0.0 & 0.0	0.79	1	0.872
	0.6 & 0.0	0.81	0.98	0.87	0.2 & 0.0	0.81	0.98	0.872

Table 6.18: Results of the Novelty detection model using ASR and low level features for *all topics* over Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
EdgeHist	0.3	0.73	0.98	0.809	0.3	0.73	0.98	0.809
	0.1	0.71	1	0.808	0.2	0.71	1	0.809
	0.4	0.76	0.86	0.78	0	0.71	1	0.808
HSVColour	0	0.7	0.93	0.764	0	0.7	0.93	0.764
	0.1	0.91	0.18	0.284	0.1	0.91	0.18	0.284
	0.2	0.94	0.13	0.218	0.2	0.94	0.13	0.218
HSVColour_CannyEd	0.2	0.72	0.93	0.771	0.2	0.72	0.93	0.771
	0.1	0.71	0.93	0.767	0.1	0.71	0.93	0.767
	0	0.7	0.93	0.764	0	0.7	0.93	0.764
HSVColour_CannyEd_ Texture	0	0.7	0.95	0.773	0	0.7	0.95	0.773
	0.1	0.7	0.94	0.77	0.1	0.7	0.94	0.77
	0.3	0.71	0.88	0.753	0.3	0.71	0.88	0.753
HSVColour_Texture	0	0.7	0.94	0.764	0	0.7	0.94	0.764
	0.1	0.8	0.34	0.444	0.1	0.8	0.34	0.444
	0.2	0.84	0.26	0.369	0.2	0.84	0.26	0.369
CannyEd	0.2	0.73	0.99	0.816	0.2	0.73	0.99	0.816
	0.1	0.72	1	0.811	0.1	0.72	1	0.811
	0	0.71	1	0.808	0	0.71	1	0.808
CannyEd_Texture	0.2	0.72	1	0.811	0.2	0.72	1	0.811
	0.1	0.71	1	0.808	0.1	0.71	1	0.808
	0.3	0.72	0.93	0.792	0	0.71	1	0.808
Texture	0	0.71	1	0.808	0	0.71	1	0.808
	0.1	0.82	0.33	0.443	0.1	0.82	0.33	0.443
	0.2	0.86	0.23	0.339	0.2	0.86	0.23	0.339
ColourStruc	0.2	0.74	0.99	0.822	0.2	0.74	0.99	0.822
	0.3	0.76	0.91	0.815	0.3	0.76	0.91	0.815
	0	0.71	1	0.808	0.1	0.72	1	0.815
ColourStruc_EdgeHist	0.6	0.74	0.99	0.822	0.6	0.74	0.99	0.822
	0.5	0.73	1	0.816	0.7	0.75	0.95	0.822
	0.4	0.72	1	0.814	0.8	0.77	0.89	0.809

Table 6.19: Results of the Novelty detection model using low level features for *all topics* over Collection_2

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_HSVColour_ CannyEd	0.0 & 0.2	0.71	0.91	0.763	0.0 & 0.2	0.71	0.91	0.763
	0.0 & 0.0	0.7	0.92	0.756	0.2 & 0.2	0.72	0.89	0.756
	0.2 & 0.0	0.71	0.89	0.752	0.6 & 0.2	0.72	0.88	0.752
ASR_HSVColour	0.0 & 0.0	0.7	0.92	0.756	0.2 & 0.0	0.71	0.89	0.752
	0.4 & 0.0	0.71	0.89	0.752	0.8 & 0.0	0.71	0.88	0.747
	0.6 & 0.0	0.71	0.89	0.749	0.6 & 0.0	0.71	0.89	0.749
ASR_HSVColour_Texture	0.0 & 0.0	0.7	0.92	0.757	0.8 & 0.0	0.71	0.88	0.747
	0.4 & 0.0	0.71	0.89	0.753	0.2 & 0.0	0.71	0.89	0.753
	0.6 & 0.0	0.71	0.89	0.75	0.6 & 0.0	0.71	0.89	0.75
ASR_ColourStruc_ Texture_CannyEd	0.0 & 0.0	0.7	0.93	0.765	0.2 & 0.2	0.71	0.9	0.758
	0.0 & 0.2	0.7	0.93	0.763	0.0 & 0.2	0.7	0.93	0.763
	0.2 & 0.0	0.71	0.9	0.761	0.2 & 0.0	0.71	0.9	0.761
ASR_CannyEd	**0.0 & 0.2**	**0.73**	**0.97**	**0.808**	**0.0 & 0.2**	**0.73**	**0.97**	**0.808**
	0.0 & 0.0	0.71	0.98	0.8	0.2 & 0.2	0.73	0.95	0.8
	0.6 & 0.2	0.73	0.94	0.797	0.6 & 0.2	0.73	0.94	0.797
ASR_Texture	**0.0 & 0.0**	**0.71**	**0.98**	0.8	**0.8 & 0.0**	**0.72**	**0.95**	**0.791**
	0.2 & 0.0	0.72	0.96	0.796	0.2 & 0.0	0.72	0.96	0.796
	0.6 & 0.0	0.72	0.95	0.793	0.6 & 0.0	0.72	0.95	0.793
ASR_Texture_CannyEd	0.0 & 0.2	0.72	0.98	0.804	0.0 & 0.2	0.72	0.98	0.804
	0.0 & 0.0	0.71	0.98	0.8	0.6 & 0.2	0.72	0.95	0.796
	0.2 & 0.2	0.72	0.95	0.799	0.2 & 0.2	0.72	0.95	0.799
ASR_ColourStruc	0.0 & 0.2	0.78	0.6	0.592	0.0 & 0.2	0.78	0.6	0.592
	0.0 & 0.0	0.77	0.61	0.586	0.0 & 0.0	0.77	0.61	0.586
	0.2 & 0.2	0.79	0.58	0.584	0.6 & 0.2	0.79	0.58	0.583
ASR_ColourStruc_ EdgeHist	0.0 & 0.6	0.78	0.61	0.592	0.0 & 0.8	0.81	0.55	0.58
	0.0 & 0.4	0.77	0.61	0.591	0.0 & 0.6	0.78	0.61	0.592
	0.0 & 0.2	0.77	0.61	0.586	0.2 & 0.8	0.81	0.54	0.57
ASR_EdgeHist	0.0 & 0.2	0.77	0.61	0.588	0.0 & 0.2	0.77	0.61	0.588
	0.0 & 0.0	0.77	0.61	0.586	0.0 & 0.0	0.77	0.61	0.586
	0.2 & 0.0	0.78	0.59	0.58	0.6 & 0.0	0.78	0.59	0.579

Table 6.20: Results of the Novelty detection model using ASR and low level features for *all topics* over Collection_2

6.4.1 "General Object" Topic Category

Tables 6.21, 6.22, 6.23 and 6.24 display both the optimal and unbiased F-measure values of the novelty run using low level features over all topics in the "General Object" topic category from Collection_1 and Collection_2 respectively.

From Tables 6.21 and 6.22 we observe the low level feature runs over the "General Object" category over Collection_1. We observe that the runs using edge histograms, "EdgeHist" and Canny edge, "CannyEd" perform well in both precision and F-measure on the baseline performance figures. Both of the runs "ColourStruc", colour structure, and "ColourStruc_EdgeHist", the combination of colour structure and edge histograms, achieve the highest Fscore of 0.975 an improvement of 5.3% on the baseline performance. When ASR is combined with low level features the highest performing run is "ASR_ColourStruc_EdgeHist" which is the combination of ASR with colour structure and edge histogram features. However this combination offers no improvement on the novelty performance of the original "ColourStruc_EdgeHist" run. We observe that the combination of ASR with low level feature runs in general over Collection_1 either has no effect or degrades the novelty performance of each run.

Once again run "ColourStruc_EdgeHist", the combination of colour structure and edge histogram achieves the highest novelty performance of all the runs for the "General Object" category over Collection_2 achieving an Fscore of 0.898 and a corresponding precision value of 0.835, an improvement of 2.5% and 6% respectively on the baseline figures (see Table 6.23). We observe that colour structure, edge histograms and Canny edge provide an improvement on the baseline performance figures. From Table 6.24, we can see that combining ASR with each of the low level features reduces the performance of the novelty detection models over the "General object" category. The highest performing Fscore during the combination of ASR and low level features is the "ASR_CannyEd" which achieves an Fscore of 0.886.

We conclude that the best consistently performing low level feature novelty detection over both collections for the "General Object" category is the combination of colour structure and edge Histogram, "ColourStruc_EdgeHist".

6.4.2 "Other" Topic Category

Tables 6.25, 6.26, 6.27 and 6.28 display both the optimal and unbiased F-measure values of the novelty run using low level features over all topics in the "Other" topic category from Collection_1 and Collection_2 respectively.

From Tables 6.25 and 6.26 we observe the performance of the low level features over the "Other" category within Collection_1. The two highest performing runs "ColourStruc_EdgeHist" and "CannyEd" achieve an Fscore of 0.925 with a corresponding precision value of 0.864 an increase of 1.5% and 2.4% on each of the baseline performance figures respectively. Colour Structure, "ColourStruc" also performs well over the "Other" category achieving an Fscore of 0.922, a 1.2% improvement on the baseline Fscore. From Table 6.25 we see that the combination of ASR with low level feature runs, increases the performance of most of the runs, however it degrades the performance of the original Canny edge run, colour structure run and the run which utilises the combination of colour structure and edge histogram. From Table 6.27 we observe that the highest performing run over Collection_2 is colour structure, "ColourStruc" achieving an Fscore of 0.866 an increase of 2.2% on the baseline while the corresponding precision value of 0.790 is 5.2% above the baseline precision figure. The "ColourStruc_EdgeHist" run achieved an Fscore of 0.858 an improvement of 1.3% while "CannyEd" achieved an Fscore of 0.857 an improvement of 1.2% on the baseline figure. If we look at the combination of ASR with each of the low level features over Collection_2, Table 6.28, we observe a decrease in the novelty performance over all runs.

We conclude that a number of low level features perform well at detecting novel shots within the "Other" category over both collections including colour struc-

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
EdgeHist	0.3	0.895	1.000	0.941	0.3	0.895	1.000	0.941
	0.4	0.928	0.963	0.939	0.0	0.865	1.000	0.926
	0	0.865	1.000	0.926	0.4	0.928	0.963	0.939
HSVColour	0	0.865	1.000	0.926	0.0	0.865	1.000	0.926
	0.1	0.945	0.332	0.486	0.1	0.945	0.332	0.486
	0.2	1.000	0.258	0.408	0.3	1.000	0.258	0.408
HSVColour_CannyEd	0	0.865	1.000	0.926	0.2	0.865	1.000	0.926
	0.3	0.878	0.982	0.925	0.3	0.878	0.982	0.925
	0.4	0.950	0.903	0.923	0.0	0.865	1.000	0.926
HSVColour_CannyEd_Texture	0	0.865	1.000	0.926	0.0	0.865	1.000	0.926
	0.4	0.862	0.982	0.916	0.4	0.862	0.982	0.916
	0.5	0.895	0.903	0.893	0.5	0.895	0.903	0.893
HSVColour_Texture	0	0.865	1.000	0.926	0.0	0.865	1.000	0.926
	0.1	0.892	0.675	0.760	0.1	0.892	0.675	0.760
	0.2	0.958	0.565	0.696	0.2	0.958	0.565	0.696
CannyEd	0.3	0.925	0.963	0.940	0.2	0.882	1.000	0.934
	0.2	0.882	1.000	0.934	0.3	0.925	0.963	0.940
	0	0.865	1.000	0.926	0.0	0.865	1.000	0.926
CannyEd_Texture	0	0.865	1.000	0.926	0.0	0.865	1.000	0.926
	0.4	0.862	0.982	0.916	0.3	0.865	1.000	0.926
	0.5	0.895	0.903	0.893	0.4	0.862	0.982	0.916
Texture	0	0.865	1.000	0.926	0.0	0.865	1.000	0.926
	0.1	0.880	0.638	0.730	0.1	0.880	0.638	0.730
	0.2	0.958	0.510	0.663	0.2	0.958	0.510	0.663
ColourStruc	0.3	0.953	1.000	0.975	0.3	0.953	1.000	0.975
	0.4	0.953	0.987	0.967	0.2	0.878	1.000	0.932
	0.2	0.878	1.000	0.932	0.1	0.865	1.000	0.926
ColourStruc_EdgeHist	0.8	0.953	1.000	0.975	0.7	0.912	1.000	0.951
	0.7	0.912	1.000	0.951	0.6	0.878	1.000	0.932
	1	0.970	0.940	0.950	0.5	0.878	1.000	0.932

Table 6.21: Results of the Novelty detection model using low level features for the "General Object" topic category over Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_HSVColour_ CannyEd	0.0 & 0.0	0.865	1.000	0.926	0.0 & 0.2	0.865	1.000	0.926
	0.0 & 0.4	0.950	0.903	0.923	0.0 & 0.0	0.865	1.000	0.926
	0.2 & 0.0	0.898	0.945	0.913	1.4 & 0.0	0.897	0.930	0.905
ASR_HSVColour	0.0 & 0.0	0.865	1.000	0.926	0.0 & 0.0	0.865	1.000	0.926
	0.2 & 0.0	0.898	0.945	0.913	1.4 & 0.0	0.897	0.930	0.905
	1.4 & 0.0	0.897	0.930	0.905	0.2 & 0.0	0.898	0.945	0.913
ASR_HSVColour_Textur	0.0 & 0.0	0.865	1.000	0.926	0.0 & 0.0	0.865	1.000	0.926
	0.2 & 0.0	0.898	0.945	0.913	1.4 & 0.0	0.897	0.930	0.905
	1.4 & 0.0	0.897	0.930	0.905	0.2 & 0.0	0.898	0.945	0.913
ASR_ColourStruc_ Texture_CannyEd	0.0 & 0.0	0.865	1.000	0.926	0.0 & 0.0	0.865	1.000	0.926
	0.0 & 0.4	0.862	0.982	0.916	1.4 & 0.0	0.897	0.930	0.905
	0.2 & 0.0	0.898	0.945	0.913	0.2 & 0.0	0.898	0.945	0.913
ASR_CannyEd	0.0 & 0.2	0.882	1.000	0.934	0.0 & 0.2	0.882	1.000	0.934
	0.0 & 0.0	0.865	1.000	0.926	0.0 & 0.0	0.865	1.000	0.926
	0.2 & 0.2	0.915	0.945	0.922	1.4 & 0.2	0.913	0.930	0.914
ASR_Texture	0.0 & 0.0	0.865	1.000	0.926	0.0 & 0.0	0.865	1.000	0.926
	0.2 & 0.0	0.898	0.945	0.913	1.4 & 0.0	0.897	0.930	0.905
	1.4 & 0.0	0.897	0.930	0.905	0.2 & 0.0	0.898	0.945	0.913
ASR_Texture_CannyEd	0.0 & 0.0	0.865	1.000	0.926	0.0 & 0.0	0.865	1.000	0.926
	0.0 & 0.4	0.862	0.982	0.916	1.4 & 0.0	0.897	0.930	0.905
	0.2 & 0.0	0.898	0.945	0.913	0.2 & 0.0	0.898	0.945	0.913
ASR_ColourStruc	0.0 & 0.4	0.953	0.987	0.967	0.0 & 0.2	0.878	1.000	0.932
	0.2 & 0.4	0.953	0.930	0.935	0.0 & 0.0	0.865	1.000	0.926
	0.0 & 0.2	0.878	1.000	0.932	0.0 & 0.4	0.953	0.987	0.967
ASR_ColourStruc_ EdgeHist	0.0 & 0.8	0.953	1.000	0.975	0.0 & 0.6	0.878	1.000	0.932
	0.0 & 1.0	0.970	0.940	0.950	0.0 & 0.0	0.865	1.000	0.926
	0.2 & 0.8	0.953	0.945	0.944	0.0 & 0.8	0.953	1.000	0.975
ASR_EdgeHist	0.0 & 0.4	0.928	0.963	0.939	0.0 & 0.4	0.928	0.963	0.939
	0.0 & 0.0	0.865	1.000	0.926	0.0 & 0.0	0.865	1.000	0.926
	0.2 & 0.0	0.898	0.945	0.913	1.4 & 0.4	0.928	0.893	0.900

Table 6.22: Results of the Novelty detection model using ASR and low level features for the "General Object" topic category over Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
EdgeHist	0.3	0.820	0.985	0.888	0.3	0.820	0.985	0.888
	0	0.787	1.000	0.876	0.4	0.828	0.897	0.846
	0.4	0.828	0.897	0.846	0.0	0.787	1.000	0.876
HSVColour	0	0.787	1.000	0.876	0.0	0.787	1.000	0.876
	0.1	0.912	0.252	0.377	0.1	0.912	0.252	0.377
	0.3	1.000	0.178	0.301	0.2	0.945	0.163	0.277
HSVColour_CannyEd	0.2	0.802	1.000	0.884	0.0	0.787	1.000	0.876
	0.3	0.810	0.975	0.879	0.3	0.810	0.975	0.879
	0	0.787	1.000	0.876	0.4	0.793	0.787	0.781
HSVColour_CannyEd_Texture	0	0.787	1.000	0.876	0.0	0.787	1.000	0.876
	0.4	0.798	0.925	0.852	0.4	0.798	0.925	0.852
	0.5	0.763	0.737	0.733	0.5	0.763	0.737	0.733
HSVColour_Texture	0	0.787	1.000	0.876	0.0	0.787	1.000	0.876
	0.1	0.820	0.468	0.576	0.1	0.820	0.468	0.576
	0.2	0.875	0.370	0.507	0.2	0.875	0.370	0.507
CannyEd	0.2	0.813	1.000	0.891	0.3	0.830	0.943	0.876
	0.3	0.830	0.943	0.876	0.2	0.813	1.000	0.891
	0	0.787	1.000	0.876	0.0	0.787	1.000	0.876
CannyEd_Texture	0	0.787	1.000	0.876	0.0	0.787	1.000	0.876
	0.3	0.783	0.985	0.868	0.4	0.793	0.900	0.838
	0.4	0.793	0.900	0.838	0.5	0.763	0.737	0.733
Texture	0	0.787	1.000	0.876	0.0	0.787	1.000	0.876
	0.1	0.807	0.445	0.552	0.1	0.807	0.445	0.552
	0.2	0.917	0.340	0.482	0.2	0.917	0.340	0.482
ColourStruc	0.3	0.833	0.980	0.896	0.3	0.833	0.980	0.896
	0.2	0.810	1.000	0.889	0.4	0.797	0.838	0.812
	0.1	0.802	1.000	0.884	0.2	0.810	1.000	0.889
ColourStruc_EdgeHist	0.7	0.835	0.985	0.898	0.8	0.825	0.943	0.876
	0.6	0.815	1.000	0.893	0.7	0.835	0.985	0.898
	0.5	0.810	1.000	0.889	1.0	0.827	0.778	0.787

Table 6.23: Results of the Novelty detection model using low level features for the "General Object" topic category over Collection_2

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_HSVColour_ CannyEd	0.0 & 0.2	0.800	0.990	0.880	0.0 & 0.0	0.785	0.990	0.871
	0.0 & 0.0	0.785	0.990	0.871	0.0 & 0.4	0.793	0.787	0.781
	1.4 & 0.0	0.822	0.932	0.863	0.2 & 0.0	0.815	0.932	0.858
ASR_HSVColour	0.0 & 0.0	0.785	0.990	0.871	0.0 & 0.0	0.785	0.990	0.871
	1.4 & 0.0	0.822	0.932	0.863	0.2 & 0.0	0.815	0.932	0.858
	0.2 & 0.0	0.815	0.932	0.858	1.4 & 0.0	0.822	0.932	0.863
ASR_HSVColour_Texture	0.0 & 0.0	0.785	0.990	0.871	0.0 & 0.0	0.785	0.990	0.871
	1.4 & 0.0	0.822	0.932	0.863	0.2 & 0.0	0.815	0.932	0.858
	0.2 & 0.0	0.815	0.932	0.858	1.4 & 0.0	0.822	0.932	0.863
ASR_ColourStruc_ Texture_CannyEd	0.0 & 0.0	0.785	0.990	0.871	0.0 & 0.0	0.785	0.990	0.871
	1.4 & 0.0	0.822	0.932	0.863	0.0 & 0.4	0.785	0.990	0.871
	0.2 & 0.0	0.815	0.932	0.858	0.2 & 0.0	0.815	0.932	0.858
ASR_CannyEd	0.0 & 0.2	0.812	0.990	0.886	0.0 & 0.2	0.812	0.990	0.886
	0.0 & 0.0	0.785	0.990	0.871	0.0 & 0.0	0.785	0.990	0.871
	1.4 & 0.2	0.833	0.932	0.869	0.2 & 0.2	0.827	0.932	0.864
ASR_Texture	0.0 & 0.0	0.785	0.990	0.871	0.0 & 0.0	0.785	0.990	0.871
	1.4 & 0.0	0.822	0.932	0.863	0.2 & 0.0	0.815	0.932	0.858
	0.2 & 0.0	0.815	0.932	0.858	1.4 & 0.0	0.822	0.932	0.863
ASR_Texture_CannyEd	0.0 & 0.0	0.785	0.990	0.871	0.0 & 0.0	0.785	0.990	0.871
	1.4 & 0.0	0.822	0.932	0.863	0.0 & 0.4	0.793	0.900	0.838
	0.2 & 0.0	0.815	0.932	0.858	0.2 & 0.0	0.815	0.932	0.858
ASR_ColourStruc	0.0 & 0.2	0.870	0.472	0.519	0.0 & 0.4	0.900	0.450	0.511
	0.0 & 0.0	0.862	0.472	0.514	0.2 & 0.4	0.895	0.392	0.477
	0.0 & 0.4	0.900	0.450	0.511	0.0 & 0.2	0.870	0.472	0.519
ASR_ColourStruc_ EdgeHist	0.0 & 0.6	0.870	0.472	0.519	0.0 & 0.8	0.890	0.450	0.505
	0.0 & 0.0	0.862	0.472	0.514	0.0 & 1.0	0.890	0.408	0.481
	0.0 & 0.8	0.890	0.450	0.505	0.2 & 0.8	0.885	0.392	0.471
ASR_EdgeHist	0.0 & 0.4	0.903	0.472	0.528	0.0 & 0.4	0.903	0.472	0.528
	0.0 & 0.0	0.862	0.472	0.514	0.0 & 0.0	0.862	0.472	0.514
	1.4 & 0.4	0.908	0.413	0.499	0.2 & 0.0	0.888	0.413	0.489

Table 6.24: Results of the Novelty detection model using ASR and low level features for the "General Object" topic category over Collection_2

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
EdgeHist	0.2	0.853	1.000	0.918	0.2	0.853	1.000	0.918
	0	0.844	1.000	0.911	0.0	0.844	1.000	0.911
	0.3	0.867	0.963	0.909	0.3	0.867	0.963	0.909
HSVColour	0	0.806	0.876	0.805	0.0	0.806	0.876	0.805
	0.1	0.940	0.149	0.251	0.1	0.940	0.149	0.251
	0.2	1.000	0.111	0.197	0.2	1.000	0.111	0.197
HSVColour_CannyEd	0.1	0.826	0.876	0.818	0.2	0.833	0.861	0.815
	0.2	0.833	0.861	0.815	0.1	0.826	0.876	0.818
	0	0.806	0.876	0.805	0.0	0.806	0.876	0.805
HSVColour_CannyEd_Texture	0.1	0.840	0.924	0.863	0.2	0.846	0.903	0.855
	0	0.836	0.924	0.860	0.1	0.840	0.924	0.863
	0.2	0.846	0.903	0.855	0.0	0.836	0.924	0.860
HSVColour_Texture	0	0.829	0.904	0.841	0.0	0.829	0.904	0.841
	0.1	0.941	0.364	0.511	0.1	0.941	0.364	0.511
	0.2	0.937	0.254	0.389	0.2	0.937	0.254	0.389
CannyEd	0.1	0.864	1.000	0.925	0.1	0.864	1.000	0.925
	0.2	0.877	0.981	0.923	0.2	0.877	0.981	0.923
	0	0.844	1.000	0.911	0.0	0.844	1.000	0.911
CannyEd_Texture	0.2	0.866	0.986	0.918	0.2	0.866	0.986	0.918
	0.1	0.849	1.000	0.914	0.1	0.849	1.000	0.914
	0	0.844	1.000	0.911	0.0	0.844	1.000	0.911
Texture	0	0.844	1.000	0.911	0.0	0.844	1.000	0.911
	0.1	0.919	0.361	0.510	0.1	0.919	0.361	0.510
	0.2	0.916	0.223	0.355	0.2	0.916	0.223	0.355
ColourStruc	0.2	0.869	0.989	0.922	0.2	0.869	0.989	0.922
	0.1	0.859	1.000	0.921	0.1	0.859	1.000	0.921
	0	0.844	1.000	0.911	0.3	0.896	0.927	0.909
ColourStruc_EdgeHist	0.4	0.864	1.000	0.925	0.5	0.866	0.996	0.924
	0.5	0.866	0.996	0.924	0.4	0.864	1.000	0.925
	0.6	0.879	0.974	0.921	0.6	0.879	0.974	0.921

Table 6.25: Results of the Novelty detection model using low level features for the "Other" topic category over Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_HSVColour_ CannyEd	0.0 & 0.2	0.833	0.861	0.815	0.0 & 0.2	0.833	0.861	0.815
	0.2 & 0.2	0.833	0.854	0.813	0.2 & 0.2	0.833	0.854	0.813
	0.2 & 0.0	0.814	0.869	0.809	0.8 & 0.2	0.833	0.854	0.813
ASR_HSVColour	0.2 & 0.0	0.814	0.869	0.809	0.2 & 0.0	0.814	0.869	0.809
	0.0 & 0.0	0.806	0.876	0.805	0.8 & 0.0	0.814	0.869	0.809
	1.2 & 0.0	0.821	0.844	0.802	0.0 & 0.0	0.806	0.876	0.805
ASR_HSVColour_Textur	0.2 & 0.0	0.837	0.897	0.845	0.2 & 0.0	0.837	0.897	0.845
	0.0 & 0.0	0.829	0.904	0.841	0.8 & 0.0	0.837	0.897	0.845
	1.2 & 0.0	0.844	0.873	0.839	0.0 & 0.0	0.829	0.904	0.841
ASR_ColourStruc_ Texture_CannyEd	0.2 & 0.0	0.844	0.917	0.864	0.0 & 0.2	0.846	0.903	0.855
	0.0 & 0.0	0.836	0.924	0.860	0.2 & 0.2	0.853	0.897	0.856
	1.2 & 0.0	0.851	0.893	0.858	0.8 & 0.2	0.853	0.897	0.856
ASR_CannyEd	0.0 & 0.2	0.877	0.981	0.923	0.0 & 0.2	0.877	0.981	0.923
	0.2 & 0.2	0.879	0.976	0.920	0.2 & 0.0	0.853	0.993	0.915
	0.2 & 0.0	0.853	0.993	0.915	0.2 & 0.2	0.879	0.976	0.920
ASR_Texture	0.2 & 0.0	0.853	0.993	0.915	0.2 & 0.0	0.853	0.993	0.915
	0.0 & 0.0	0.844	1.000	0.911	0.8 & 0.0	0.853	0.993	0.915
	1.2 & 0.0	0.860	0.969	0.909	0.0 & 0.0	0.844	1.000	0.911
ASR_Texture_CannyEd	0.2 & 0.2	0.871	0.979	0.919	0.0 & 0.2	0.866	0.986	0.918
	0.0 & 0.2	0.866	0.986	0.918	0.2 & 0.2	0.871	0.979	0.919
	0.2 & 0.0	0.853	0.993	0.915	1.0 & 0.2	0.871	0.957	0.910
ASR_ColourStruc	0.0 & 0.2	0.855	0.987	0.914	0.0 & 0.2	0.855	0.987	0.914
	0.2 & 0.2	0.857	0.980	0.911	0.2 & 0.0	0.837	0.992	0.906
	0.2 & 0.0	0.837	0.992	0.906	0.2 & 0.2	0.857	0.980	0.911
ASR_ColourStruc_ EdgeHist	0.0 & 0.4	0.850	1.000	0.917	0.0 & 0.4	0.850	1.000	0.917
	0.2 & 0.4	0.850	0.992	0.914	0.2 & 0.4	0.850	0.992	0.914
	0.0 & 0.6	0.867	0.970	0.912	1.0 & 0.4	0.850	0.975	0.907
ASR_EdgeHist	0.0 & 0.2	0.837	1.000	0.908	0.0 & 0.2	0.837	1.000	0.908
	0.2 & 0.0	0.837	0.992	0.906	0.2 & 0.0	0.837	0.992	0.906
	1.2 & 0.0	0.845	0.972	0.903	0.0 & 0.0	0.827	1.000	0.901

Table 6.26: Results of the Novelty detection model using ASR and low level features for the "Other" topic category over Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
EdgeHist	0.2	0.760	1.000	0.852	0.2	0.760	1.000	0.852
	0	0.751	1.000	0.847	0.0	0.751	1.000	0.847
	0.3	0.773	0.957	0.840	0.3	0.773	0.957	0.840
HSVColour	0	0.756	0.907	0.799	0.0	0.756	0.907	0.799
	0.1	0.946	0.119	0.207	0.1	0.946	0.119	0.207
	0.2	0.971	0.083	0.150	0.2	0.971	0.083	0.150
HSVColour_CannyEd	0.2	0.780	0.897	0.810	0.1	0.773	0.906	0.809
	0.1	0.773	0.906	0.809	0.2	0.780	0.897	0.810
	0	0.756	0.907	0.799	0.0	0.756	0.907	0.799
HSVColour_CannyEd_Texture	0.2	0.751	0.917	0.812	0.1	0.740	0.921	0.807
	0.1	0.740	0.921	0.807	0.0	0.731	0.921	0.802
	0	0.731	0.921	0.802	0.2	0.751	0.917	0.812
HSVColour_Texture	0	0.744	0.914	0.802	0.0	0.744	0.914	0.802
	0.1	0.800	0.254	0.375	0.1	0.800	0.254	0.375
	0.2	0.844	0.180	0.290	0.2	0.844	0.180	0.290
CannyEd	0.1	0.769	0.999	0.857	0.1	0.769	0.999	0.857
	0.2	0.774	0.979	0.852	0.2	0.774	0.979	0.852
	0	0.751	1.000	0.847	0.0	0.751	1.000	0.847
CannyEd_Texture	0.2	0.771	0.993	0.857	0.2	0.771	0.993	0.857
	0.1	0.756	1.000	0.850	0.1	0.756	1.000	0.850
	0	0.751	1.000	0.847	0.0	0.751	1.000	0.847
Texture	0	0.751	1.000	0.847	0.0	0.751	1.000	0.847
	0.1	0.833	0.259	0.384	0.1	0.833	0.259	0.384
	0.2	0.831	0.150	0.249	0.2	0.831	0.150	0.249
ColourStruc	0.2	0.790	0.987	0.866	0.2	0.790	0.987	0.866
	0.1	0.767	1.000	0.857	0.1	0.767	1.000	0.857
	0.3	0.817	0.913	0.848	0.0	0.751	1.000	0.847
ColourStruc_EdgeHist	0.5	0.771	0.996	0.858	0.4	0.769	0.999	0.857
	0.4	0.769	0.999	0.857	0.5	0.771	0.996	0.858
	0.6	0.780	0.970	0.853	0.6	0.780	0.970	0.853

Table 6.27: Results of the Novelty detection model using low level features for the "Other" topic category over Collection_2

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_HSVColour_ CannyEd	0.0 & 0.2	0.773	0.870	0.797	0.0 & 0.2	0.773	0.870	0.797
	0.2 & 0.2	0.773	0.864	0.794	0.2 & 0.2	0.773	0.864	0.794
	0.8 & 0.2	0.773	0.861	0.794	0.2 & 0.0	0.757	0.873	0.788
ASR_HSVColour	0.2 & 0.0	0.757	0.873	0.788	0.2 & 0.0	0.757	0.873	0.788
	0.8 & 0.0	0.757	0.870	0.788	0.0 & 0.0	0.749	0.880	0.786
	0.0 & 0.0	0.749	0.880	0.786	1.2 & 0.0	0.766	0.850	0.786
ASR_HSVColour_Texture	0.2 & 0.0	0.746	0.880	0.791	0.2 & 0.0	0.746	0.880	0.791
	0.8 & 0.0	0.746	0.877	0.790	0.0 & 0.0	0.737	0.887	0.789
	0.0 & 0.0	0.737	0.887	0.789	1.2 & 0.0	0.754	0.857	0.788
ASR_ColourStruc_ Texture_CannyEd	0.0 & 0.2	0.733	0.894	0.793	0.2 & 0.0	0.733	0.887	0.791
	0.2 & 0.2	0.737	0.887	0.793	0.0 & 0.0	0.724	0.894	0.789
	0.8 & 0.2	0.737	0.884	0.792	1.2 & 0.0	0.741	0.864	0.788
ASR_CannyEd	0.0 & 0.2	0.767	0.951	0.838	0.0 & 0.2	0.767	0.951	0.838
	0.2 & 0.0	0.753	0.966	0.836	0.2 & 0.2	0.767	0.946	0.836
	0.2 & 0.2	0.767	0.946	0.836	0.2 & 0.0	0.753	0.966	0.836
ASR_Texture	0.2 & 0.0	0.753	0.966	0.836	0.2 & 0.0	0.753	0.966	0.836
	0.8 & 0.0	0.753	0.963	0.835	0.0 & 0.0	0.744	0.973	0.834
	0.0 & 0.0	0.744	0.973	0.834	1.2 & 0.0	0.761	0.943	0.833
ASR_Texture_CannyEd	0.0 & 0.2	0.764	0.966	0.843	0.2 & 0.2	0.769	0.959	0.843
	0.2 & 0.2	0.769	0.959	0.843	0.0 & 0.2	0.764	0.966	0.843
	1.0 & 0.2	0.769	0.956	0.842	0.2 & 0.0	0.753	0.966	0.836
ASR_ColourStruc	0.0 & 0.2	0.822	0.543	0.590	0.0 & 0.2	0.822	0.543	0.590
	0.2 & 0.0	0.808	0.547	0.588	0.2 & 0.2	0.822	0.538	0.587
	0.2 & 0.2	0.822	0.538	0.587	0.2 & 0.0	0.808	0.547	0.588
ASR_ColourStruc_ EdgeHist	0.0 & 0.4	0.822	0.553	0.599	0.0 & 0.4	0.822	0.553	0.599
	0.2 & 0.4	0.822	0.547	0.596	0.2 & 0.4	0.822	0.547	0.596
	1.0 & 0.4	0.818	0.538	0.591	0.0 & 0.6	0.825	0.532	0.583
ASR_EdgeHist	0.0 & 0.2	0.808	0.553	0.592	0.0 & 0.2	0.808	0.553	0.592
	0.2 & 0.0	0.808	0.547	0.588	0.2 & 0.0	0.808	0.547	0.588
	0.0 & 0.0	0.797	0.553	0.585	1.2 & 0.0	0.810	0.528	0.580

Table 6.28: Results of the Novelty detection model using ASR and low level features for the "Other" topic category over Collection_2

ture and Canny edge. We note that the combination of colour structure and edge histogram low level features, consistently perform well over both collections.

6.4.3 "People" Topic Category

Tables 6.29, 6.30, 6.31 and 6.32 display both the optimal and unbiased F-measure values of the novelty run using low level features over all topics in the "People" topic category from Collection_1 and Collection_2 respectively.

From Tables 6.29 and 6.30 we observe the performances of the low level features over the "People" category within Collection_1. We observe that both of the edge feature runs, namely the edge histogram run, "EdgeHist" and Canny edge "CannyEd" perform well during the detection of novel shots, providing an improvement of 3.9% and 1.9% on the baseline novelty performance figures respectively. colour achieved a slightly lower novelty performance than edge features with colour structure, "ColourStruc" achieved an Fscore of 0.840 an improvement of 1.7%, while HSV colour, "HSVColour" performed similar to the baseline, returning all documents as novel. A combination of HSV colour and Canny edge, "HSVColour_CannyEd", increased the performance of each of the individual novelty detection runs. Texture does not aid in the detection of novel shots in the "People" category over Collection_1. The combination of colour structure and edge histogram, "ColourStruc_EdgeHist" achieved the highest performing Fscore of 0.873 an improvement of 6% on the baseline figure of 0.823, while precision, 0.800, achieved an improvement of 10.3% on the baseline precision figure of 0.725. Combining Canny edge and texture achieves an Fscore above the baseline, however it is performing lower than Canny edge on its own. If we look at the combination of ASR with each of the low level features in Table 6.30 we note that many runs achieve an increase in novelty performance however, the combination of ASR with the highest performing low-level feature run "ColourStruc_EdgeHist" decreases the overall novelty performance of the original run.

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
EdgeHist	0.4	0.812	0.945	0.862	0.4	0.812	0.945	0.862
	0.3	0.740	0.992	0.832	0.3	0.740	0.992	0.832
	0	0.725	1.000	0.823	0.0	0.725	1.000	0.823
HSVColour	0	0.725	1.000	0.823	0.0	0.725	1.000	0.823
	0.1	0.883	0.205	0.333	0.1	0.883	0.205	0.333
	0.2	1.000	0.150	0.263	0.2	1.000	0.150	0.263
HSVColour_CannyEd	0.3	0.755	1.000	0.845	0.4	0.773	0.930	0.829
	0.2	0.733	1.000	0.830	0.2	0.733	1.000	0.830
	0.4	0.773	0.930	0.829	0.1	0.725	1.000	0.823
HSVColour_CannyEd_Texture	0.3	0.738	1.000	0.832	0.3	0.738	1.000	0.832
	0.2	0.732	1.000	0.828	0.2	0.732	1.000	0.828
	0	0.725	1.000	0.823	0.0	0.725	1.000	0.823
HSVColour_Texture	0	0.725	1.000	0.823	0.0	0.725	1.000	0.823
	0.1	0.812	0.455	0.575	0.1	0.812	0.455	0.575
	0.2	0.868	0.293	0.435	0.2	0.868	0.293	0.435
CannyEd	0.3	0.763	0.980	0.842	0.2	0.748	1.000	0.840
	0.2	0.748	1.000	0.840	0.1	0.727	1.000	0.824
	0.1	0.727	1.000	0.824	0.0	0.725	1.000	0.823
CannyEd_Texture	0.3	0.738	1.000	0.832	0.3	0.738	1.000	0.832
	0.2	0.732	1.000	0.828	0.2	0.732	1.000	0.828
	0	0.725	1.000	0.823	0.0	0.725	1.000	0.823
Texture	0	0.725	1.000	0.823	0.0	0.725	1.000	0.823
	0.1	0.798	0.370	0.498	0.1	0.798	0.370	0.498
	0.2	0.830	0.215	0.332	0.2	0.830	0.215	0.332
ColourStruc	0.3	0.800	0.913	0.840	0.2	0.753	0.987	0.840
	0.2	0.753	0.987	0.840	0.3	0.800	0.913	0.840
	0.1	0.733	1.000	0.830	0.1	0.733	1.000	0.830
ColourStruc_EdgeHist	0.7	0.800	0.987	0.873	0.8	0.830	0.928	0.866
	0.9	0.877	0.877	0.872	0.7	0.800	0.987	0.873
	0.8	0.830	0.928	0.866	0.6	0.760	0.992	0.847

Table 6.29: Results of the Novelty detection model using low level features for the "People" topic category over Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_HSVColour_ CannyEd	0.2 & 0.2	0.747	0.995	0.836	0.0 & 0.4	0.773	0.930	0.829
	0.2 & 0.4	0.790	0.925	0.834	0.0 & 0.2	0.733	1.000	0.830
	0.0 & 0.2	0.733	1.000	0.830	0.2 & 0.4	0.790	0.925	0.834
ASR_HSVColour	0.2 & 0.0	0.738	0.995	0.828	0.0 & 0.0	0.725	1.000	0.823
	0.0 & 0.0	0.725	1.000	0.823	0.2 & 0.0	0.738	0.995	0.828
	0.6 & 0.0	0.737	0.980	0.821	0.6 & 0.0	0.737	0.980	0.821
ASR_HSVColour_Texture	0.2 & 0.0	0.738	0.995	0.828	0.0 & 0.0	0.725	1.000	0.823
	0.0 & 0.0	0.725	1.000	0.823	0.2 & 0.0	0.738	0.995	0.828
	0.6 & 0.0	0.737	0.980	0.821	0.6 & 0.0	0.737	0.980	0.821
ASR_ColourStruc_ Texture_CannyEd	0.2 & 0.2	0.745	0.995	0.834	0.0 & 0.4	0.733	0.968	0.818
	0.2 & 0.0	0.738	0.995	0.828	0.0 & 0.0	0.725	1.000	0.823
	0.0 & 0.2	0.732	1.000	0.828	0.2 & 0.4	0.748	0.963	0.824
ASR_CannyEd	0.2 & 0.2	0.762	0.995	0.846	0.0 & 0.2	0.748	1.000	0.840
	0.0 & 0.2	0.748	1.000	0.840	0.2 & 0.2	0.762	0.995	0.846
	0.6 & 0.2	0.760	0.980	0.839	0.0 & 0.0	0.725	1.000	0.823
ASR_Texture	0.2 & 0.0	0.738	0.995	0.828	0.0 & 0.0	0.725	1.000	0.823
	0.0 & 0.0	0.725	1.000	0.823	0.2 & 0.0	0.738	0.995	0.828
	0.6 & 0.0	0.737	0.980	0.821	0.6 & 0.0	0.737	0.980	0.821
ASR_Texture_CannyEd	0.2 & 0.2	0.745	0.995	0.834	0.0 & 0.2	0.732	1.000	0.828
	0.2 & 0.0	0.738	0.995	0.828	0.0 & 0.0	0.725	1.000	0.823
	0.0 & 0.2	0.732	1.000	0.828	0.2 & 0.2	0.745	0.995	0.834
ASR_ColourStruc	0.2 & 0.2	0.770	0.982	0.846	0.0 & 0.2	0.753	0.987	0.840
	0.0 & 0.2	0.753	0.987	0.840	0.2 & 0.2	0.770	0.982	0.846
	0.6 & 0.2	0.767	0.967	0.838	0.6 & 0.2	0.767	0.967	0.838
ASR_ColourStruc_ EdgeHist	0.0 & 0.8	0.830	0.928	0.866	0.0 & 0.8	0.830	0.928	0.866
	0.2 & 0.8	0.835	0.923	0.866	0.2 & 0.8	0.835	0.923	0.866
	0.6 & 0.8	0.833	0.912	0.860	0.6 & 0.8	0.833	0.912	0.860
ASR_EdgeHist	0.4 & 0.4	0.815	0.945	0.864	0.2 & 0.4	0.812	0.945	0.862
	0.0 & 0.4	0.812	0.945	0.862	0.0 & 0.4	0.812	0.945	0.862
	0.6 & 0.4	0.813	0.935	0.858	0.6 & 0.4	0.813	0.935	0.858

Table 6.30: Results of the Novelty detection model using ASR and low level features for the "People" topic category over Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
EdgeHist	0.4	0.653	0.903	0.743	0.4	0.653	0.903	0.743
	0.3	0.603	0.983	0.728	0.3	0.603	0.983	0.728
	0	0.598	1.000	0.725	0.0	0.598	1.000	0.725
HSVColour	0	0.598	1.000	0.725	0.0	0.598	1.000	0.725
	0.1	0.823	0.178	0.289	0.1	0.823	0.178	0.289
	0.2	0.850	0.117	0.204	0.2	0.850	0.117	0.204
HSVColour_CannyEd	0.4	0.667	0.872	0.733	0.3	0.608	0.958	0.723
	0.2	0.605	1.000	0.732	0.2	0.605	1.000	0.732
	0.1	0.600	1.000	0.727	0.4	0.667	0.872	0.733
HSVColour_CannyEd_Texture	0.3	0.605	0.987	0.729	0.3	0.605	0.987	0.729
	0.2	0.602	1.000	0.728	0.2	0.602	1.000	0.728
	0	0.598	1.000	0.725	0.0	0.598	1.000	0.725
HSVColour_Texture	0	0.598	1.000	0.725	0.0	0.598	1.000	0.725
	0.1	0.660	0.332	0.433	0.1	0.660	0.332	0.433
	0.2	0.710	0.223	0.335	0.2	0.710	0.223	0.335
CannyEd	0.2	0.608	0.995	0.735	0.3	0.625	0.927	0.722
	0.1	0.602	1.000	0.728	0.2	0.608	0.995	0.735
	0	0.598	1.000	0.725	0.1	0.602	1.000	0.728
CannyEd_Texture	0.3	0.605	0.987	0.729	0.3	0.605	0.987	0.729
	0.2	0.602	1.000	0.728	0.2	0.602	1.000	0.728
	0	0.598	1.000	0.725	0.0	0.598	1.000	0.725
Texture	0	0.598	1.000	0.725	0.0	0.598	1.000	0.725
	0.1	0.718	0.287	0.404	0.1	0.718	0.287	0.404
	0.2	0.728	0.178	0.276	0.2	0.728	0.178	0.276
ColourStruc	0.2	0.627	0.985	0.747	0.3	0.665	0.883	0.741
	0.3	0.665	0.883	0.741	0.2	0.627	0.985	0.747
	0.1	0.605	1.000	0.732	0.1	0.605	1.000	0.732
ColourStruc_EdgeHist	0.8	0.695	0.913	0.776	0.7	0.655	0.958	0.765
	0.7	0.655	0.958	0.765	0.9	0.725	0.807	0.752
	0.6	0.633	0.985	0.753	0.8	0.695	0.913	0.776

Table 6.31: Results of the Novelty detection model using low level features for the "People" topic category over Collection_2

	Optimised Results				Unbiased Results			
Runs	**T**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_HSVColour_ CannyEd	0.0 & 0.4	0.667	0.868	0.731	0.2 & 0.2	0.607	0.960	0.723
	0.0 & 0.2	0.605	0.997	0.730	0.2 & 0.4	0.673	0.848	0.729
	0.2 & 0.4	0.673	0.848	0.729	0.0 & 0.2	0.605	0.997	0.730
ASR_HSVColour	0.0 & 0.0	0.598	0.997	0.723	0.2 & 0.0	0.602	0.960	0.717
	0.2 & 0.0	0.602	0.960	0.717	0.0 & 0.0	0.598	0.997	0.723
	0.6 & 0.0	0.597	0.935	0.705	0.6 & 0.0	0.597	0.935	0.705
ASR_HSVColour_Texture	0.0 & 0.0	0.598	0.997	0.723	0.2 & 0.0	0.602	0.960	0.717
	0.2 & 0.0	0.602	0.960	0.717	0.0 & 0.0	0.598	0.997	0.723
	0.6 & 0.0	0.597	0.935	0.705	0.6 & 0.0	0.597	0.935	0.705
ASR_ColourStruc_ Texture_CannyEd	0.0 & 0.4	0.605	0.983	0.727	0.2 & 0.2	0.602	0.960	0.717
	0.0 & 0.0	0.598	0.997	0.723	0.2 & 0.0	0.602	0.960	0.717
	0.2 & 0.4	0.607	0.947	0.720	0.0 & 0.2	0.598	0.997	0.723
ASR_CannyEd	0.0 & 0.2	0.608	0.992	0.733	0.2 & 0.2	0.610	0.955	0.726
	0.2 & 0.2	0.610	0.955	0.726	0.0 & 0.2	0.608	0.992	0.733
	0.0 & 0.0	0.598	0.997	0.723	0.6 & 0.2	0.605	0.930	0.714
ASR_Texture	0.0 & 0.0	0.598	0.997	0.723	0.2 & 0.0	0.602	0.960	0.717
	0.2 & 0.0	0.602	0.960	0.717	0.0 & 0.0	0.598	0.997	0.723
	0.6 & 0.0	0.597	0.935	0.705	0.6 & 0.0	0.597	0.935	0.705
ASR_Texture_CannyEd	0.0 & 0.2	0.602	0.997	0.727	0.2 & 0.2	0.603	0.960	0.720
	0.0 & 0.0	0.598	0.997	0.723	0.2 & 0.0	0.602	0.960	0.717
	0.2 & 0.2	0.603	0.960	0.720	0.0 & 0.2	0.602	0.997	0.727
ASR_ColourStruc	0.0 & 0.2	0.687	0.555	0.526	0.2 & 0.2	0.697	0.535	0.522
	0.2 & 0.2	0.697	0.535	0.522	0.0 & 0.2	0.687	0.555	0.526
	0.6 & 0.2	0.697	0.530	0.518	0.6 & 0.2	0.697	0.530	0.518
ASR_ColourStruc_ EdgeHist	0.0 & 0.8	0.740	0.520	0.537	0.0 & 0.8	0.740	0.520	0.537
	0.2 & 0.8	0.745	0.517	0.536	0.2 & 0.8	0.745	0.517	0.536
	0.6 & 0.8	0.743	0.513	0.534	0.6 & 0.8	0.743	0.513	0.534
ASR_EdgeHist	0.2 & 0.4	0.710	0.520	0.522	0.4 & 0.4	0.710	0.520	0.522
	0.0 & 0.4	0.707	0.520	0.521	0.0 & 0.4	0.707	0.520	0.521
	0.6 & 0.4	0.710	0.513	0.517	0.6 & 0.4	0.710	0.513	0.517

Table 6.32: Results of the Novelty detection model using ASR and low level features for the "People" topic category over Collection_2

From Tables 6.31 and 6.32 we observe the performance of each of the low level feature runs on the "People" category over Collection_2. We observe that both edge feature runs once again perform above the baseline novelty performance figures. colour structure "ColourStruc", however, performs higher than edge features over this collection providing an improvement of 3.0% on the baseline respectively, while HSV colour "HSVColour" obtained an Fscore similar to the baseline performance. This characteristic was observed over Collection_1 also and as a result we conclude that HSV colour does not aid in identifying novel shots in the "People" category. Once again we observe that texture does not perform above the baseline in Collection_2 and as a result we conclude that it does not aid in identifying novel shots for the "People" category in general. The combination of HSV colour and Canny edge, "HSVColour_CannyEd" performs lower than the novelty performance of the Canny edge run on its own. This is caused by the low performance of the HSV colour run over this category. The combination of colour structure and edge histogram, "ColourStruc_EdgeHist" however improves on the performance of both of the colour structure and edge histogram runs separately, becoming the highest performing run over Collection_2, achieving an Fscore of 0.776 and a corresponding precision value of 0.695, an improvement of 7% and 16.22% on the baseline novelty performance figures. If we look at the combination of ASR with each of the low level features and over the "People" category within Collection_2, Table 6.32 we observe that ASR degrades the performance of each of the original runs for the detection of novel shots.

In conclusion, each of the individual, colour structure, edge histograms and Canny edge low level features perform well on the "People" category over both collections, however the combination of colour structure and edge histograms outperform all other runs in the detection of novel shots within the "People" category.

6.4.4 "Specific Object" Topic Category

Tables 6.33, 6.34, 6.35 and 6.36 display both the optimal and unbiased F-measure values of the novelty run using low level features over all topics in the "Specific Object" category from Collection_1 and Collection_2 respectively.

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
EdgeHist	0.5	1.000	1.000	1.000	0.5	1.000	1.000	1.000
	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.6	1.000	0.500	0.667	0.6	1.000	0.500	0.667
HSVColour	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.1	1.000	0.500	0.667	0.1	1.000	0.500	0.667
	0.7	1.000	0.250	0.400	0.7	1.000	0.250	0.400
HSVColour_CannyEd	0.6	1.000	1.000	1.000	0.6	1.000	1.000	1.000
	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.7	1.000	0.500	0.667	0.7	1.000	0.500	0.667
HSVColour_CannyEd_Texture	0.6	1.000	1.000	1.000	0.6	1.000	1.000	1.000
	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.8	1.000	0.750	0.857	0.8	1.000	0.750	0.857
HSVColour_Texture	0.1	1.000	1.000	1.000	0.1	1.000	1.000	1.000
	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.2	1.000	0.750	0.857	0.2	1.000	0.750	0.857
CannyEd	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.5	0.750	0.750	0.750	0.5	0.750	0.750	0.750
	0.6	1.000	0.500	0.667	0.6	1.000	0.500	0.667
CannyEd_Texture	0.6	1.000	1.000	1.000	0.6	1.000	1.000	1.000
	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.8	1.000	0.750	0.857	0.8	1.000	0.750	0.857
Texture	0.1	1.000	1.000	1.000	0.1	1.000	1.000	1.000
	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.2	1.000	0.750	0.857	0.2	1.000	0.750	0.857
ColourStruc	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.5	0.750	0.750	0.750	0.5	0.750	0.750	0.750
	0.6	0.670	0.500	0.571	0.6	0.670	0.500	0.571
ColourStruc_EdgeHist	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	1.2	0.670	0.500	0.571	1.2	0.670	0.500	0.571
	1.5	1.000	0.250	0.400	1.5	1.000	0.250	0.400

Table 6.33: Results of the Novelty detection model using low level features for the "Specific Object" topic category over Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_HSVColour_ CannyEd	0.0 &0.6	1.000	1.000	1.000	0.0 &0.6	1.000	1.000	1.000
	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	10.0 &0.6	1.000	0.750	0.857	10.0 &0.6	1.000	0.750	0.857
ASR_HSVColour	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	10.6 & 0.0	1.000	0.750	0.857	10.6 & 0.0	1.000	0.750	0.857
	10.0 & 0.0	0.750	0.750	0.750	10.0 & 0.0	0.750	0.750	0.750
ASR_HSVColour_Texture	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	0.0 & 0.2	1.000	0.750	0.857	0.0 & 0.2	1.000	0.750	0.857
	10.0 & 0.0	0.750	0.750	0.750	10.0 & 0.0	0.750	0.750	0.750
ASR_ColourStruc_ Texture_CannyEd	0.0 &0.6	1.000	1.000	1.000	1.0 & 0.0	0.800	1.000	0.889
	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	0.0 & 0.8	1.000	0.750	0.857	0.0 & 0.0	0.800	1.000	0.889
ASR_CannyEd	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	10.6 & 0.0	1.000	0.750	0.857	10.0 & 0.0	0.750	0.750	0.750
	10.0 & 0.0	0.750	0.750	0.750	0.0 &0.6	1.000	0.500	0.667
ASR_Texture	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	0.0 & 0.2	1.000	0.750	0.857	0.0 & 0.2	1.000	0.750	0.857
	10.0 & 0.0	0.750	0.750	0.750	10.0 & 0.0	0.750	0.750	0.750
ASR_Texture_CannyEd	0.0 &0.6	1.000	1.000	1.000	0.0 &0.6	1.000	1.000	1.000
	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	0.0 & 0.8	1.000	0.750	0.857	0.0 & 0.8	1.000	0.750	0.857
ASR_ColourStruc	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	0.0 & 0.6	0.670	0.500	0.571	0.0 & 0.6	0.670	0.500	0.571
	0.0 & 1.0	1.000	0.250	0.400	0.0 & 1.0	1.000	0.250	0.400
ASR_ColourStruc_ EdgeHist	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	0.0 & 1.2	0.670	0.500	0.571	0.0 & 1.2	0.670	0.500	0.571
	0.0 & 1.6	1.000	0.250	0.400	0.0 & 1.6	1.000	0.250	0.400
ASR_EdgeHist	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	0.0 & 0.6	1.000	0.500	0.667	0.0 & 0.6	1.000	0.500	0.667
	0.0 & 0.8	1.000	0.250	0.400	0.0 & 0.8	1.000	0.250	0.400

Table 6.34: Results of the Novelty detection model using ASR and low level features for the "Specific Object" topic category over Collection_1

From Table 6.33 we observe that HSV colour, colour structure and Canny edge features and a combination of colour structure and edge histograms do not achieve any improvement upon the baseline figures. This is surprising as the combination of colour structure and edge histograms have performed well on all category topics so far. We observe that the edge histogram and texture runs, "EdgeHist" and "Texture" respectively achieve an Fscore of 1 and a corresponding precision value of 1, an improvement of 12.5% on the baseline Fscore. The run which combines HSV colour and Canny edge, "HSVColour_CannyEd" achieves an Fscore of 1 and corresponding precision value of 1. This is an improvement on the runs which use only these low level feature individually. Similarly the runs which utilise a combination of HSV colour, Canny edge and texture, "HSVColour_CannyEd_Texture", HSV colour and Texture, "HSV-Colour_Texture" and Canny edge and texture "CannyEd_Texture" all achieve Fscores of 1 and precision values of 1. Runs achieving Fscores of 1 and a corresponding precision value of 1, are performing similar to the highest performing manual assessor. If we look at the combination of ASR with each of the low level feature runs in Table 6.34, we observe that ASR has either no effect or reduces the performance of each of the novelty runs over the "Specific object" category within Collection_1. If we look at the low level runs over the "Specific Object" category within Collection_2, Table 6.35 and 6.36 we notice a similar trend for each of the low level feature runs.

This leads us to conclude that novelty models using texture and edge histograms are good low level feature resources in identifying novel shots within the "Specific Object" category. We also conclude that a combination of various features such as texture, Canny edge and HSV colour, perform well in identifying novel shots within the "Specific Object" category.

6.4.5 "Sports" Topic Category

Tables 6.37, 6.38, 6.39 and 6.40 display both the optimal and unbiased F-measure values of the novelty run using low level features over all topics in the

"Sports" topic category from Collection_1 and Collection_2 respectively.

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
EdgeHist	0.5	1.000	1.000	1.000	0.5	1.000	1.000	1.000
	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.6	1.000	0.500	0.667	0.6	1.000	0.500	0.667
HSVColour	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.1	1.000	0.500	0.667	0.1	1.000	0.500	0.667
	0.7	1.000	0.250	0.400	0.7	1.000	0.250	0.400
HSVColour_CannyEd	0.6	1.000	1.000	1.000	0.6	1.000	1.000	1.000
	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.7	1.000	0.500	0.667	0.7	1.000	0.500	0.667
HSVColour_CannyEd_Texture	0.6	1.000	1.000	1.000	0.6	1.000	1.000	1.000
	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.8	1.000	0.750	0.857	0.8	1.000	0.750	0.857
HSVColour_Texture	0.1	1.000	1.000	1.000	0.1	1.000	1.000	1.000
	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.2	1.000	0.750	0.857	0.2	1.000	0.750	0.857
CannyEd	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.5	0.750	0.750	0.750	0.5	0.750	0.750	0.750
	0.6	1.000	0.500	0.667	0.6	1.000	0.500	0.667
CannyEd_Texture	0.6	1.000	1.000	1.000	0.6	1.000	1.000	1.000
	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.8	1.000	0.750	0.857	0.8	1.000	0.750	0.857
Texture	0.1	1.000	1.000	1.000	0.1	1.000	1.000	1.000
	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.2	1.000	0.750	0.857	0.2	1.000	0.750	0.857
ColourStruc	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.5	0.750	0.750	0.750	0.5	0.750	0.750	0.750
	0.6	0.670	0.500	0.571	0.6	0.670	0.500	0.571
ColourStruc_EdgeHist	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	1.2	0.670	0.500	0.571	1.2	0.670	0.500	0.571
	1.5	1.000	0.250	0.400	1.5	1.000	0.250	0.400

Table 6.35: Results of the Novelty detection model using low level features for the "Specific Object" topic category over Collection_2

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_HSVColour_ CannyEd	0.0 & 0.6	1.000	1.000	1.000	0.0 &0.6	1.000	1.000	1.000
	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	10.0 & 0.6	1.000	0.750	0.857	10.0 &0.6	1.000	0.750	0.857
ASR_HSVColour	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	10.6 & 0.0	1.000	0.750	0.857	10.6 & 0.0	1.000	0.750	0.857
	10.0 & 0.0	0.750	0.750	0.750	10.0 & 0.0	0.750	0.750	0.750
ASR_HSVColour_Textur	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	0.0 & 0.2	1.000	0.750	0.857	0.0 & 0.2	1.000	0.750	0.857
	10.0 & 0.0	0.750	0.750	0.750	10.0 & 0.0	0.750	0.750	0.750
ASR_ColourStruc_ Texture_CannyEd	0.0 & 0.7	1.000	1.000	1.000	2.0 & 0.0	0.800	1.000	0.889
	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	0.0 & 0.9	1.000	0.750	0.857	0.0 & 0.0	0.800	1.000	0.889
ASR_CannyEd	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	10.0 & 0.0	0.750	0.750	0.750	0.0 & 0.0	0.800	1.000	0.889
	0.0 & 0.6	1.000	0.500	0.667	10.0 & 0.0	0.750	0.750	0.750
ASR_Texture	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	0.0 & 0.2	1.000	0.750	0.857	0.0 & 0.2	1.000	0.750	0.857
	10.0 & 0.0	0.750	0.750	0.750	10.0 & 0.0	0.750	0.750	0.750
ASR_Texture_CannyEd	0.0 & 0.6	1.000	1.000	1.000	0.0 &0.6	1.000	1.000	1.000
	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	0.0 & 0.8	1.000	0.750	0.857	0.0 & 0.8	1.000	0.750	0.857
ASR_ColourStruc	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	0.0 & 0.6	0.670	0.500	0.571	0.0 & 0.6	0.670	0.500	0.571
	0.0 & 1.0	1.000	0.250	0.400	0.0 & 1.0	1.000	0.250	0.400
ASR_ColourStruc_ EdgeHist	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	0.0 & 1.2	0.670	0.500	0.571	0.0 & 1.2	0.670	0.500	0.571
	0.0 & 1.6	1.000	0.250	0.400	0.0 & 1.6	1.000	0.250	0.400
ASR_EdgeHist	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	0.0 & 0.6	1.000	0.500	0.667	0.0 & 0.6	1.000	0.500	0.667
	0.0 & 0.8	1.000	0.250	0.400	0.0 & 0.8	1.000	0.250	0.400

Table 6.36: Results of the Novelty detection model using ASR and low level features for the "Specific Object" topic category over Collection_2

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T**	**Prec**	**Recall**	**Fscore**
EdgeHist	0.3	0.660	1.000	0.771	0.0	0.657	1.000	0.768
	0	0.657	1.000	0.768	0.3	0.660	1.000	0.771
	0.4	0.697	0.797	0.739	0.4	0.697	0.797	0.739
HSVColour	0	0.680	0.737	0.563	0.0	0.680	0.737	0.563
	0.1	1.000	0.087	0.155	0.1	1.000	0.087	0.155
					0.2	1.000	0.087	0.155
HSVColour_CannyEd	0.2	0.687	0.737	0.570	0.0	0.680	0.737	0.563
	0	0.680	0.737	0.563	0.3	0.670	0.503	0.496
	0.4	0.777	0.447	0.521	0.6	0.917	0.253	0.353
HSVColour_CannyEd_Texture	0	0.680	0.787	0.621	0.0	0.680	0.787	0.621
	0.2	0.683	0.763	0.600	0.1	0.680	0.763	0.594
	0.1	0.680	0.763	0.594	0.2	0.683	0.763	0.600
HSVColour_Texture	0	0.680	0.737	0.563	0.0	0.680	0.737	0.563
	0.1	1.000	0.140	0.226	0.1	1.000	0.140	0.226
	0.3	1.000	0.087	0.155	0.3	1.000	0.087	0.155
CannyEd	0.2	0.663	1.000	0.775	0.0	0.657	1.000	0.768
	0	0.657	1.000	0.768	0.3	0.650	0.767	0.702
	0.4	0.777	0.687	0.724	0.5	0.850	0.587	0.660
CannyEd_Texture	0.2	0.660	1.000	0.774	0.0	0.657	1.000	0.768
	0	0.657	1.000	0.768	0.2	0.660	1.000	0.774
	0.4	0.783	0.710	0.743	0.3	0.640	0.780	0.700
Texture	0	0.657	1.000	0.768	0.0	0.657	1.000	0.768
	0.1	1.000	0.187	0.299	0.1	1.000	0.187	0.299
	0.2	1.000	0.163	0.265	0.2	1.000	0.163	0.265
ColourStruc	0.3	0.730	0.900	0.801	0.0	0.657	1.000	0.768
	0.2	0.667	0.983	0.776	0.2	0.667	0.983	0.776
	0.4	0.803	0.800	0.772	0.3	0.730	0.900	0.801
ColourStruc_EdgeHist	0.9	0.790	0.843	0.802	0.6	0.663	0.993	0.773
	0.8	0.723	0.877	0.786	0.5	0.657	1.000	0.770
	0.6	0.663	0.993	0.773	0.0	0.657	1.000	0.768

Table 6.37: Results of the Novelty detection model using low level features for the "Sport" topic category over Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_HSVColour_ CannyEd	1.8 & 0.2	0.693	0.720	0.575	0.8 & 0.2	0.687	0.720	0.569
	1.6 & 0.2	0.693	0.720	0.574	0.8 & 0.0	0.680	0.720	0.562
	2.0 & 0.2	0.693	0.713	0.573	0.2 & 0.2	0.687	0.730	0.571
ASR_HSVColour	1.8 & 0.0	0.687	0.720	0.569	0.8 & 0.0	0.680	0.720	0.562
	1.6 & 0.0	0.687	0.720	0.568	0.2 & 0.0	0.680	0.730	0.564
	2.0 & 0.0	0.687	0.713	0.567	1.0 & 0.0	0.683	0.720	0.564
ASR_HSVColour_Texture	1.8 & 0.0	0.687	0.720	0.569	0.8 & 0.0	0.680	0.720	0.562
	1.6 & 0.0	0.687	0.720	0.568	0.2 & 0.0	0.680	0.730	0.564
	2.0 & 0.0	0.687	0.713	0.567	1.0 & 0.0	0.683	0.720	0.564
ASR_ColourStruc_ Texture_CannyEd	1.8 & 0.0	0.687	0.770	0.627	0.8 & 0.0	0.680	0.770	0.619
	1.6 & 0.0	0.687	0.770	0.625	0.2 & 0.0	0.680	0.780	0.621
	2.0 & 0.0	0.687	0.763	0.625	1.0 & 0.0	0.683	0.770	0.622
ASR_CannyEd	1.8 & 0.2	0.670	0.983	0.780	0.8 & 0.2	0.663	0.983	0.773
	1.6 & 0.2	0.670	0.983	0.778	0.8 & 0.0	0.657	0.983	0.766
	2.0 & 0.2	0.670	0.977	0.778	0.2 & 0.2	0.663	0.993	0.775
ASR_Texture	1.8 & 0.0	0.663	0.983	0.773	0.8 & 0.0	0.657	0.983	0.766
	1.6 & 0.0	0.663	0.983	0.772	0.2 & 0.0	0.657	0.993	0.768
	2.0 & 0.0	0.663	0.977	0.771	1.0 & 0.0	0.660	0.983	0.768
ASR_Texture_CannyEd	1.8 & 0.2	0.670	0.983	0.778	0.8 & 0.0	0.657	0.983	0.766
	1.6 & 0.2	0.667	0.983	0.777	1.0 & 0.2	0.663	0.983	0.775
	2.0 & 0.2	0.667	0.977	0.776	0.2 & 0.0	0.657	0.993	0.768
ASR_ColourStruc	0.2 & 0.2	0.738	0.983	0.826	1.4 & 0.2	0.740	0.965	0.820
	0.0 & 0.2	0.738	0.988	0.826	1.2 & 0.2	0.740	0.965	0.819
	0.8 & 0.2	0.738	0.978	0.823	0.8 & 0.0	0.730	0.988	0.818
ASR_ColourStruc_ EdgeHist	0.0 & 0.8	0.793	0.908	0.839	1.2 & 0.6	0.738	0.970	0.817
	0.2 & 0.8	0.793	0.900	0.837	0.8 & 0.6	0.735	0.983	0.822
	1.0 & 0.8	0.793	0.888	0.830	0.2 & 0.6	0.735	0.988	0.824
ASR_EdgeHist	0.6 & 0.0	0.733	0.995	0.821	1.2 & 0.0	0.733	0.975	0.814
	0.2 & 0.0	0.730	0.995	0.819	1.4 & 0.0	0.733	0.975	0.815
	0.0 & 0.0	0.730	1.000	0.819	0.8 & 0.0	0.730	0.988	0.818

Table 6.38: Results of the Novelty detection model using ASR and low level features for the "Sport" topic category over Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
EdgeHist	0	0.667	1.000	0.718	0.3	0.663	0.980	0.715
	0.3	0.663	0.980	0.715	0.0	0.667	1.000	0.718
	0.4	0.670	0.773	0.665	0.4	0.670	0.773	0.665
HSVColour	0	0.600	0.717	0.492	0.0	0.600	0.717	0.492
	0.1	1.000	0.073	0.136	0.1	1.000	0.073	0.136
					0.1	1.000	0.073	0.136
HSVColour_CannyEd	0	0.600	0.717	0.492	0.2	0.600	0.697	0.492
	0.3	0.597	0.457	0.433	0.0	0.600	0.717	0.492
	0.6	0.933	0.270	0.406	0.4	0.603	0.323	0.388
HSVColour_CannyEd_Texture	0	0.643	0.770	0.556	0.0	0.643	0.770	0.556
	0.1	0.627	0.743	0.526	0.2	0.627	0.723	0.525
	0.2	0.627	0.723	0.525	0.1	0.627	0.743	0.526
HSVColour_Texture	0	0.600	0.717	0.492	0.0	0.600	0.717	0.492
	0.1	1.000	0.103	0.180	0.1	1.000	0.103	0.180
	0.3	1.000	0.073	0.136	0.3	1.000	0.073	0.136
CannyEd	0	0.667	1.000	0.718	0.2	0.667	0.980	0.718
	0.3	0.663	0.740	0.660	0.0	0.667	1.000	0.718
	0.5	0.720	0.530	0.593	0.4	0.663	0.557	0.587
CannyEd_Texture	0	0.667	1.000	0.718	0.2	0.667	0.980	0.717
	0.2	0.667	0.980	0.717	0.0	0.667	1.000	0.718
	0.3	0.663	0.787	0.676	0.4	0.657	0.557	0.581
Texture	0	0.667	1.000	0.718	0.0	0.667	1.000	0.718
	0.1	1.000	0.153	0.257	0.1	1.000	0.153	0.257
	0.2	1.000	0.127	0.221	0.2	1.000	0.127	0.221
ColourStruc	0	0.667	1.000	0.718	0.3	0.673	0.797	0.703
	0.2	0.667	0.960	0.717	0.2	0.667	0.960	0.717
	0.3	0.673	0.797	0.703	0.4	0.693	0.700	0.690
ColourStruc_EdgeHist	0.6	0.670	1.000	0.724	0.9	0.683	0.720	0.687
	0.5	0.667	1.000	0.719	0.8	0.660	0.720	0.662
	0	0.667	1.000	0.718	0.6	0.670	1.000	0.724

Table 6.39: Results of the Novelty detection model using low level features for the "Sport" topic category over Collection_2

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_HSVColour_ CannyEd	0.8 & 0.2	0.603	0.670	0.482	1.8 & 0.2	0.597	0.630	0.475
	0.8 & 0.0	0.600	0.690	0.482	1.6 & 0.2	0.597	0.630	0.474
	0.2 & 0.2	0.600	0.670	0.481	2.0 & 0.2	0.600	0.630	0.476
ASR_HSVColour	0.8 & 0.0	0.600	0.690	0.482	1.8 & 0.0	0.597	0.630	0.471
	0.2 & 0.0	0.600	0.690	0.480	1.6 & 0.0	0.597	0.630	0.470
	1.0 & 0.0	0.600	0.670	0.479	2.0 & 0.0	0.597	0.630	0.472
ASR_HSVColour_Texture	0.8 & 0.0	0.600	0.690	0.482	1.8 & 0.0	0.597	0.630	0.471
	0.2 & 0.0	0.600	0.690	0.480	1.6 & 0.0	0.597	0.630	0.470
	1.0 & 0.0	0.600	0.670	0.479	2.0 & 0.0	0.597	0.630	0.472
ASR_ColourStruc_ Texture_CannyEd	0.8 & 0.0	0.643	0.743	0.546	1.8 & 0.0	0.640	0.683	0.536
	0.2 & 0.0	0.643	0.743	0.545	1.6 & 0.0	0.640	0.683	0.535
	1.0 & 0.0	0.643	0.723	0.543	2.0 & 0.0	0.640	0.683	0.537
ASR_CannyEd	0.8 & 0.2	0.670	0.953	0.708	1.8 & 0.2	0.663	0.913	0.701
	0.8 & 0.0	0.667	0.973	0.708	1.6 & 0.2	0.663	0.913	0.700
	0.2 & 0.2	0.667	0.953	0.707	2.0 & 0.2	0.667	0.913	0.702
ASR_Texture	0.8 & 0.0	0.667	0.973	0.708	1.8 & 0.0	0.663	0.913	0.697
	0.2 & 0.0	0.667	0.973	0.707	1.6 & 0.0	0.663	0.913	0.696
	1.0 & 0.0	0.667	0.953	0.705	2.0 & 0.0	0.663	0.913	0.698
ASR_Texture_CannyEd	0.8 & 0.0	0.667	0.973	0.708	1.8 & 0.2	0.663	0.913	0.699
	1.0 & 0.2	0.667	0.953	0.707	1.6 & 0.2	0.667	0.933	0.704
	0.2 & 0.0	0.667	0.973	0.707	2.0 & 0.2	0.663	0.913	0.700
ASR_ColourStruc	1.4 & 0.2	0.740	0.852	0.735	0.2 & 0.2	0.727	0.867	0.732
	1.2 & 0.2	0.740	0.852	0.735	0.0 & 0.2	0.727	0.867	0.730
	0.8 & 0.0	0.727	0.898	0.734	0.8 & 0.2	0.727	0.867	0.732
ASR_ColourStruc_ EdgeHist	1.2 & 0.6	0.743	0.883	0.740	0.0 & 0.8	0.740	0.760	0.724
	0.8 & 0.6	0.730	0.898	0.738	0.2 & 0.8	0.740	0.760	0.725
	0.2 & 0.6	0.730	0.898	0.737	1.0 & 0.8	0.740	0.760	0.725
ASR_EdgeHist	1.2 & 0.0	0.740	0.883	0.737	0.6 & 0.0	0.727	0.898	0.733
	1.4 & 0.0	0.738	0.867	0.734	0.2 & 0.0	0.727	0.898	0.733
	0.8 & 0.0	0.727	0.898	0.734	0.0 & 0.0	0.727	0.898	0.731

Table 6.40: Results of the Novelty detection model using ASR and low level features for the "Sport" topic category over Collection_2

Tables 6.37 and 6.38 allow us to analyse the performance of the low level features and their combination with and without ASR on the "Sport" category within Collection_1. We observe that many runs perform below the baseline Fscore of 0.768, including HSV colour and various combinations of runs which include this feature, including the combination of HSV colour with Canny edge and texture and this is consistent over Collection_2 (Tables 6.39). The texture run, "Texture" does not improve upon the baseline performance figures. We observe that the edge feature runs, edge histogram, "EdgeHist" and Canny edge, "CannyEd", perform novelty detection above the baseline run. We see that colour structure, "ColourStruc", aid in the detection of novel shots achieving an Fscore of 0.801 an improvement of 4.3% while obtaining a precision value of 0.730 an increase of 11.1% on the baseline. The combination of colour structure and edge histograms run, "ColourStruc_EdgeHist", achieves the highest non-ASR run Fscore of 0.802 with a corresponding precision value of 0.790. This corresponds to an improvement of 4.4% and 20.2% respectively on the baseline. This is a clear improvement on detecting novel shots within the "Sports" category when compared to runs using colour structure and edge histogram features separately. The combination of ASR and low level features displayed in Table 6.38 show that ASR improves the performance of all runs. We can clearly see that the combination of ASR with the highest performing combination of low level features, "ColourStruc_EdgeHist", achieves the highest overall novelty performance value of 0.839, an improvement of 10.2%, with a corresponding precision value of 0.793, an improvement of 17.1% on the baseline precision figure.

Once again if we observe the performances of the runs over the "Sports" category within Collection_2 (Tables 6.39 and 6.40), we see that the runs which utilise Canny edge "CannyEd" and texture "Texture" separately do not improve upon the baseline novelty performance figures. Combining these features, "CannyEd_Texture", has no effect on the novelty performance. Runs including edge histogram, "EdgeHist", and colour structure, "ColourStruc", do not improve on the baseline, however the run which uses a combination of these two features

once again achieves the highest non-ASR run Fscore of 0.724, an improvement of 0.8% on the baseline, while the corresponding precision value 0.670 is an increase of 0.4%. From Table 6.40 we observe that the combination of ASR with each of the runs, excluding colour structure, edge histogram and the combination of these feature runs, degrades the novelty detection performance. However the run which combines ASR with colour structure "ASR_ColourStruc", achieves an Fscore of 0.735, an increase of 2.4% on the baseline, while the run combining ASR with edge histogram features, achieves an Fscore of 0.737, an increase of 2.6% on the baseline figure. The highest performing run over the "Sports" category within Collection_2, the run which combines ASR with colour structure and edge histogram "ASR_ColourStruc_EdgeHist" achieved an Fscore of 0.740 an improvement of 3.1%, and a precision value of 0.743 an improvement of 11.4% on the baseline precision figure.

We conclude that combination of ASR with colour structure and edge histogram evidences is a good method of identifying novel shot within the "Sports" category as it consistently achieves the highest performance over both Collection_1 and Collection_2. This is surprising as this is the only category where ASR seems to aid the detection of novel shots.

6.4.6 Summary analysis for low level features

Low level features are the primary content extraction methods from visual content. As novelty detection within the video domain is a visual task we would expect these feature evidences to aid in the detection of novel shots from a list of shots for a topic. We have performed an exhaustive comparison of all of the low level features available for the detection of novel shots from within a list of chronologically ordered shots relevant to a topic, firstly by looking at the performance of the features over all topics in each collection and then by looking at how they perform over each of the individual topic categories as seen from Tables 6.17 through to Table 6.40 inclusive. From the findings presented above, we can conclude that colour structure, edge histograms and the combination of

both these low level features, perform well during the detection of novel shots within the video domain.

6.5 Video Novelty Model using Manually Annotated Features

In this section we investigate the performance of the novelty detection models when utilising the manually annotated concepts assigned to each shot over each topic category. We also investigate the effect on the detection of novel shots when we use a combination of text (ASR) and manually annotated concepts. Two runs namely "Concepts_Shot_By_Shot" and "ASR_Concepts_Shot_By_Shot" were explored to investigate "the shot by shot" approach to novelty detection as described in Chapter 4, while utilising manually annotated concepts and a combination of ASR and manually annotated concepts respectively. Runs "Concepts" and "ASR_Concepts" were used to explore the performance of novelty models utilising manually annotated concepts and a combination of both ASR and concepts when the model uses an accumulative history of all shots seen so far to determine the novelty of a shot. Tables 6.41 and 6.42 display the optimal novelty performance for each of the manual concept runs over all topics in both Collection_1 and Collection_2 respectively.

From Table 6.41, it is clear that the combination of ASR and concepts, "ASR_Concepts", performs better than all other runs over Collection_1 achieving an Fscore of 0.872, an improvement of 1% on the baseline performance figures. We observe from Table 6.42, that although this run only achieves the baseline performance figures for all topics in Collection_2, it is the highest performing run. We will now look at how the manual concepts perform over each of the individual topic categories.

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
Concepts	0	0.8	0.99	0.869	0	0.8	0.99	0.869
	0.1	0.88	0.66	0.736	0.1	0.88	0.66	0.736
	0.2	0.89	0.61	0.702	0.2	0.89	0.61	0.702
Concept_Shot_By_Shot	0	0.92	0.29	0.42	0	0.92	0.29	0.42
	0.1	0.93	0.25	0.375	0.1	0.93	0.25	0.375
	0.2	0.94	0.19	0.298	0.2	0.94	0.19	0.298
ASR_Concepts	0.1	0.81	1	0.881	0.1	0.81	1	0.881
	0	0.81	1	0.88	0	0.81	1	0.88
	0.2	0.81	0.99	0.877	0.2	0.81	0.99	0.877
ASR_Concepts_ Shot_by_Shot	0	0.91	0.3	0.427	0	0.91	0.3	0.427
	0.1	0.92	0.18	0.278	0.1	0.92	0.18	0.278
	0.2	0.95	0.1	0.178	0.2	0.95	0.1	0.178

Table 6.41: Results of the Novelty detection model using manually annotated concepts for *all topics* over Collection_1

6.5.1 "General Object" Topic Category

Tables 6.43 and 6.44 display both the optimal and unbiased F-measure values of the novelty run using manually annotated concepts over all topics in the "General Object" topic category from Collection_1 and Collection_2 respectively.

As we can see from Table 6.43, the manual concept run for novelty detection within the "General Object" category over Collection_1 performs below the baseline values of Collection_1 with an Fscore of 0.918. However, we can see that a combination of ASR and concepts, "ASR_Concepts", achieves an optimal Fscore of 0.946 and a corresponding precision value of 0.915, an improvement of 2.2 % and 5.8% respectively on the baseline values. In Table 6.44 we see that the manual concepts run is once again performing below the baseline for Collection_2. The combination of ASR and concepts however achieved an Fscore of 0.904, an increase of 3.2% on the baseline while the corresponding precision value of 0.867 shows an increase of 10.2% over the baseline performance. This

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
Concepts	0	0.71	0.99	0.802	0	0.71	0.99	0.802
	0.1	0.83	0.62	0.692	0.1	0.83	0.62	0.692
	0.2	0.81	0.55	0.639	0.2	0.81	0.55	0.639
Concept_Shot_By_Shot	0	0.87	0.27	0.4	0	0.87	0.27	0.4
	0.1	0.9	0.25	0.366	0.1	0.9	0.25	0.366
	0.2	0.88	0.17	0.263	0.2	0.88	0.17	0.263
ASR_Concepts	0.1	0.72	0.98	0.808	0.1	0.72	0.98	0.808
	0	0.72	0.98	0.807	0	0.72	0.98	0.807
	0.2	0.72	0.97	0.803	0.2	0.72	0.97	0.803
ASR_Concepts_ Shot_by_Shot	0	0.86	0.28	0.408	0	0.86	0.28	0.408
	0.1	0.87	0.16	0.253	0.1	0.87	0.16	0.253
	0.2	0.95	0.08	0.15	0.2	0.95	0.08	0.15

Table 6.42: Results of the Novelty detection model using manually annotated concepts for *all topics* over Collection_2

would suggest that the combination of ASR and concepts performs consistently well over the "General Object" topic categories.

6.5.2 "Other" Topic Category

Tables 6.45 and 6.46 display both the optimal and unbiased F-measure values of the novelty run using manually annotated concepts over all topics in the "Other" topic category from Collection_1 and Collection_2 respectively.

Table 6.45 shows that the run which solely utilises the manual concepts, "Concepts", achieves an Fscore similar to the baseline performance results, however the corresponding precision value of 0.849 is an improvement of 0.6% on the precision baseline result. The run which utilises a combination of ASR and concepts, "ASR_Concepts", achieves an Fscore of 0.918 and a corresponding precision value of 0.857, an improvement of 0.8% and 1.5% on the baseline performance values respectively. We can see from Table 6.46 that the perfor-

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_Concepts	1	0.915	0.987	0.946	1.9	0.915	0.987	0.946
	0	0.898	1.000	0.944	1.1	0.915	0.987	0.946
	0.5	0.898	0.987	0.937	1.7	0.915	0.987	0.946
ASR_Concepts_ Shot_by_Shot	0	0.945	0.490	0.618	0.0	0.945	0.490	0.618
	0.1	1.000	0.318	0.466	0.1	1.000	0.318	0.466
	0.2	1.000	0.188	0.309	0.2	1.000	0.188	0.309
Concept_Shot_By_Shot	0	0.945	0.490	0.618	0.1	1.000	0.433	0.590
	0.1	1.000	0.433	0.590	0.0	0.945	0.490	0.618
	0.2	1.000	0.318	0.466	0.2	1.000	0.318	0.466
Concepts	0	0.865	0.987	0.918	0.0	0.865	0.987	0.918
	0.1	0.938	0.843	0.871	0.1	0.938	0.843	0.871
	0.2	0.935	0.825	0.858	0.2	0.935	0.825	0.858

Table 6.43: Results of the Novelty detection model using manually annotated concepts for the "General Object" topic category over Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_Concepts	1.9	0.867	0.952	0.904	1.0	0.835	0.982	0.897
	1.1	0.847	0.982	0.903	0.0	0.818	0.990	0.890
	1.7	0.855	0.962	0.900	0.5	0.827	0.990	0.895
ASR_Concepts_ Shot_by_Shot	0	0.925	0.437	0.578	0.0	0.925	0.437	0.578
	0.1	0.958	0.290	0.409	0.1	0.958	0.290	0.409
	0.2	1.000	0.123	0.216	0.2	1.000	0.123	0.216
Concept_Shot_By_Shot	0.1	1.000	0.420	0.572	0.0	0.925	0.420	0.562
	0	0.925	0.420	0.562	0.1	1.000	0.420	0.572
	0.2	0.945	0.283	0.403	0.2	0.945	0.283	0.403
Concepts	0	0.790	0.983	0.872	0.0	0.790	0.983	0.872
	0.1	0.942	0.812	0.864	0.1	0.942	0.812	0.864
	0.2	0.915	0.735	0.804	0.2	0.915	0.735	0.804

Table 6.44: Results of the Novelty detection model using manually annotated concepts for the "General Object" topic category over Collection_2

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_Concepts	0.1	0.857	0.993	0.918	0.1	0.857	0.993	0.918
	0	0.851	0.993	0.914	0.0	0.851	0.993	0.914
	0.2	0.857	0.970	0.909	0.2	0.857	0.970	0.909
ASR_Concepts_ Shot_by_Shot	0	0.843	0.184	0.294	0.0	0.843	0.184	0.294
	0.1	0.823	0.104	0.178	0.1	0.823	0.104	0.178
	0.3	1.000	0.054	0.103	0.2	0.929	0.054	0.102
Concept_Shot_By_Shot	0	0.843	0.184	0.294	0.0	0.843	0.184	0.294
	0.1	0.864	0.161	0.261	0.1	0.864	0.161	0.261
	0.2	0.964	0.131	0.221	0.2	0.964	0.131	0.221
Concepts	0	0.849	0.993	0.911	0.0	0.849	0.993	0.911
	0.1	0.867	0.581	0.678	0.1	0.867	0.581	0.678
	0.2	0.880	0.520	0.637	0.2	0.880	0.520	0.637

Table 6.45: Results of the Novelty detection model using manually annotated concepts for the "Other" topic category over Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_Concepts	0.1	0.757	0.964	0.838	0.1	0.757	0.964	0.838
	0	0.753	0.964	0.835	0.0	0.753	0.964	0.835
	0.2	0.751	0.943	0.828	0.2	0.751	0.943	0.828
ASR_Concepts_ Shot_by_Shot	0	0.837	0.177	0.282	0.0	0.837	0.177	0.282
	0.1	0.810	0.087	0.156	0.1	0.810	0.087	0.156
	0.2	0.929	0.051	0.097	0.3	1.000	0.043	0.083
Concept_Shot_By_Shot	0	0.837	0.177	0.282	0.0	0.837	0.177	0.282
	0.1	0.839	0.143	0.238	0.1	0.839	0.143	0.238
	0.2	0.871	0.107	0.186	0.2	0.871	0.107	0.186
Concepts	0	0.754	0.990	0.844	0.0	0.754	0.990	0.844
	0.1	0.833	0.511	0.613	0.1	0.833	0.511	0.613
	0.2	0.847	0.467	0.585	0.2	0.847	0.467	0.585

Table 6.46: Results of the Novelty detection model using manually annotated concepts for the "Other" topic category over Collection_2

mance of this run, "ASR_Concepts", is not consistent over Collection_2, and all of the runs perform below the baseline performance values. The highest of the two runs utilises manual concepts only, "Concepts", and achieves an Fscore of 0.844. These performance results are not surprising as we have observed that the performance of runs which utilise a resource that performs below the baseline actually decrease, when combined with ASR over the collection.

6.5.3 "People" Topic Category

Tables 6.47 and 6.48 display both the optimal and unbiased F-measure values of the novelty run using manually annotated concepts over all topics in the "People" topic category from Collection_1 and Collection_2 respectively.

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_Concepts	0	0.738	1.000	0.831	0.0	0.738	1.000	0.831
	0.7	0.742	0.980	0.823	0.2	0.738	0.985	0.823
	0.2	0.738	0.985	0.823	0.5	0.738	0.980	0.821
ASR_Concepts_ Shot_by_Shot	0	0.958	0.220	0.354	0.0	0.958	0.220	0.354
	0.1	1.000	0.105	0.186	0.1	1.000	0.105	0.186
	0.2	1.000	0.057	0.109	0.2	1.000	0.057	0.109
Concept_Shot_By_Shot	0	0.958	0.220	0.354	0.0	0.958	0.220	0.354
	0.1	0.958	0.190	0.310	0.1	0.958	0.190	0.310
	0.2	0.917	0.118	0.208	0.2	0.917	0.118	0.208
Concepts	0	0.727	1.000	0.824	0.0	0.727	1.000	0.824
	0.1	0.875	0.563	0.681	0.1	0.875	0.563	0.681
	0.2	0.878	0.507	0.638	0.2	0.878	0.507	0.638

Table 6.47: Results of the Novelty detection model using manually annotated concepts for the "People" topic category Collection_1

Tables 6.47 shows that manual concepts, "Concepts", perform just above the baseline in the "People" category in Collection_1 achieving an Fscore of 0.824

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_Concepts	0	0.607	1.000	0.730	0.0	0.607	1.000	0.730
	0.2	0.603	0.975	0.721	0.7	0.612	0.937	0.712
	0.5	0.607	0.960	0.717	0.2	0.603	0.975	0.721
ASR_Concepts_ Shot_by_Shot	0	0.852	0.233	0.356	0.0	0.852	0.233	0.356
	0.1	0.888	0.125	0.207	0.1	0.888	0.125	0.207
	0.2	1.000	0.062	0.111	0.2	1.000	0.062	0.111
Concept_Shot_By_Shot	0	0.852	0.233	0.356	0.0	0.852	0.233	0.356
	0.1	0.892	0.213	0.330	0.1	0.892	0.213	0.330
	0.2	0.860	0.113	0.195	0.2	0.860	0.113	0.195
Concepts	0	0.600	1.000	0.725	0.0	0.600	1.000	0.725
	0.1	0.742	0.563	0.627	0.1	0.742	0.563	0.627
	0.2	0.692	0.493	0.559	0.2	0.692	0.493	0.559

Table 6.48: Results of the Novelty detection model using manually annotated concepts for the "People" topic category over Collection_2

an insignificant improvement of 0.1% over the baseline performance figures. However the combination of ASR and concepts, "ASR_Concepts", achieved a slightly higher Fscore of 0.831, an improvement of 1% over the the baseline. If we look at the same runs over Collection_2 in Table 6.48, we see that the individual manual concept run achieves an Fscore of 0.725. Once again this run shows an insignificant improvement over the baseline performance figures in terms of precision and no improvement in terms of Fscore for this category. The combination of ASR and concepts achieved an Fscore of 0.730 an improvement of 0.7% with a corresponding precision value of 0.607. This would suggest that the run using a combination of ASR and manually annotated concepts consistently performs well over the "People" topic category over both collections.

6.5.4 "Specific Object" Topic Category

Tables 6.49 and 6.50 display both the optimal and unbiased F-measure values of the novelty run using manually annotated concepts for all topics in the "Specific Object" topic category from Collection_1 and Collection_2 respectively.

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_Concepts	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	5.8	1.000	0.750	0.857	5.8	1.000	0.750	0.857
	2.8	0.750	0.750	0.750	2.8	0.750	0.750	0.750
ASR_Concepts_ Shot_by_Shot	0	1.000	0.500	0.667	0.0	1.000	0.500	0.667
	0.1	1.000	0.250	0.400	0.1	1.000	0.250	0.400
					1.0	1.000	0.250	0.400
Concept_Shot_By_Shot	0	1.000	0.500	0.667	0.0	1.000	0.500	0.667
	0.1	1.000	0.250	0.400	0.1	1.000	0.250	0.400
					0.2	1.000	0.250	0.400
Concepts	0.1	1.000	1.000	1.000	0.1	1.000	1.000	1.000
	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.6	1.000	0.750	0.857	0.6	1.000	0.750	0.857

Table 6.49: Results of the Novelty detection model using manually annotated concepts for the "Specific Object" topic category Collection_1

From Table 6.49 we can observe that the run utilising only manual concepts, "Concepts", for detecting novel shots over Collection_1 achieves an Fscore of 1 and a corresponding precision figure of 1, an improvement for 12.5% and 25% over the baseline figures respectively. The combination of ASR and manual concepts, "ASR_Concepts", achieves a performance figure similar to the baseline performance figures over Collection_1. The same characteristics can be observed for each run over Collection_2, see Table 6.50. This would suggest that the novelty detection models using only manually annotated concepts, "Concepts", performs well over the "Specific Object" category.

	Optimised Results				Unbiased Results			
Runs	**T**	**Prec**	**Recall**	**Fscore**	**T**	**Prec**	**Recall**	**Fscore**
ASR_Concepts	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	5.8	1.000	0.750	0.857	5.8	1.000	0.750	0.857
	2.8	0.750	0.750	0.750	2.8	0.750	0.750	0.750
ASR_Concepts_ Shot_by_Shot	0	1.000	0.500	0.667	0.0	1.000	0.500	0.667
	0.1	1.000	0.250	0.400	0.1	1.000	0.250	0.400
					1.0	1.000	0.250	0.400
Concept_Shot_By_Shot	0	1.000	0.500	0.667	0.0	1.000	0.500	0.667
	0.1	1.000	0.250	0.400	0.1	1.000	0.250	0.400
					0.2	1.000	0.250	0.400
Concepts	0.1	1.000	1.000	1.000	0.1	1.000	1.000	1.000
	0	0.800	1.000	0.889	0.0	0.800	1.000	0.889
	0.6	1.000	0.750	0.857	0.6	1.000	0.750	0.857

Table 6.50: Results of the Novelty detection model using manually annotated concepts for the "Specific Object" topic category over Collection_2

6.5.5 "Sports" Topic Category

Tables 6.51 and 6.52 display both the optimal and unbiased F-measure values of the novelty run using manually annotated concepts over all topics in the "Sports" topic category from Collection_1 and Collection_2 respectively.

From Table 6.51 and we observe that the manual concept run, "Concepts", achieves an Fscore of 0.760 which is below the baseline Fscore performance figures, however when ASR is combined with manual concepts we note that the run, "ASR_Concepts", achieves an Fscore of 0.785, an improvement of 2.2% on the baseline performance figures. The corresponding precision values of 0.717 is an improvement of 9.1% on the baseline precision value. However this performance is not consistent over Collection_2 as seen in Table 6.52, where we notice that all runs perform below the baseline performance figures. The higher of the two runs, the combination of ASR and manual concepts "ASR_Concepts", over Collection_2 achieved an Fscore of 0.706.

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_Concepts	6.5	0.717	0.867	0.785	0.0	0.660	1.000	0.771
	5	0.697	0.907	0.784	0.4	0.657	0.977	0.766
	4.8	0.693	0.907	0.782	0.3	0.660	0.983	0.770
ASR_Concepts_ Shot_by_Shot	0	0.890	0.293	0.425	0.0	0.890	0.293	0.425
	0.1	0.833	0.173	0.278	0.1	0.833	0.173	0.278
	1	1.000	0.087	0.155	1.0	1.000	0.087	0.155
Concept_ Shot_By_Shot	0	0.933	0.237	0.370	0.0	0.933	0.237	0.370
	0.1	0.850	0.210	0.333	0.1	0.850	0.210	0.333
	0.2	0.807	0.180	0.288	0.2	0.807	0.180	0.288
Concepts	0	0.670	0.943	0.760	0.0	0.670	0.943	0.760
	0.1	0.797	0.537	0.622	0.1	0.797	0.537	0.622
	0.7	0.917	0.467	0.581	0.5	0.853	0.460	0.567

Table 6.51: Results of the Novelty detection model using manually annotated concepts for the "Sports" topic category over Collection_1

6.5.6 Summary analysis for manually annotated concepts

We have performed a detailed analysis of the performance of the novelty detection models using manual concepts and highlighted the best performing runs, firstly over all topics and then over each of the topic categories, over both collections by looking at Table 6.41 to Table 6.52 inclusive. The manual content description of video using standardised concepts in the form of an ontology, is the most accurate form of content description to date of broadcast news data as seen in Chapter 4 and as a result, one would expect that models using this resource would perform well in identifying novel shots from a list of shots. However ontologies are composed of a certain number of predefined keywords and this can cause many shots to become indistinguishable from each other, in other words making them appear redundant as the findings presented above illustrated. There are a number of reasons for inconsistencies between

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_Concepts	0	0.667	0.973	0.706	6.5	0.663	0.720	0.658
	0.4	0.663	0.933	0.701	5.0	0.670	0.807	0.690
	0.3	0.663	0.933	0.700	4.8	0.670	0.807	0.688
ASR_Concepts_ Shot_by_Shot	0	0.777	0.253	0.382	0.0	0.777	0.253	0.382
	0.1	0.750	0.127	0.216	0.1	0.750	0.127	0.216
	1	1.000	0.073	0.136	1.0	1.000	0.073	0.136
Concept_ Shot_By_Shot	0	0.800	0.227	0.351	0.0	0.800	0.227	0.351
	0.1	0.800	0.200	0.313	0.1	0.800	0.200	0.313
	0.2	0.750	0.153	0.252	0.2	0.750	0.153	0.252
Concepts	0	0.667	0.943	0.689	0.0	0.667	0.943	0.689
	0.1	0.727	0.460	0.558	0.1	0.727	0.460	0.558
	0.5	0.777	0.380	0.510	0.7	0.740	0.333	0.458

Table 6.52: Results of the Novelty detection model using manually annotated concepts for the "Sports" topic category over Collection_2

Collection_1 and Collection_2 for each of the topic categories including the fact that, due to the subjectiveness of human annotation and visual perception annotated data can be inconsistent. If we consider the "Sports" category for example, which contains Topic 130 "Find shots of a hockey rink with at least one of the nets fully visible from some point of view". This topic has much redundant data, as seen from Table 6.3 and 6.4, with less than 50% of the shots considered novel. During the annotation of such data, the annotator must work with the predefined concepts in the standardised ontology to describe the shots which may contain only a certain number of suitable concepts. This can lead to almost all of the shots within the hockey topic receiving the same concept content description. During novelty detection which depends solely on these description, the majority of these shots are identified as redundant shots. The combination of ASR and manually annotated concepts, increases the number of words or concepts that are considered during the identification of a shot's

novelty value. This reduces the number of redundant shots that can occur due to the lack of concepts in the ontology and hence the over usage of specific concepts during annotation to describe a shot's content. We conclude that in this case manual concepts that fail to account for everything taking place in the image, will not perform well in novelty detection. We also observe when shots have sufficient content descriptions that manual concepts in novelty detection works well and this can be seen within the "Specific object" category.

6.6 Video Novelty Model using Automatic High Level Features

In this section we consider novelty detection models which utilise high level feature resources associated with a video sequence including: Face, Anchor, Commercial, Studio setting, Graphics, Weather, Sports, Outdoor, Person, Crowd, Road, Car, Building and Motion as proposed in Chapter 4. We investigate five different feature combinations optimally weighted for each of the specified topic categories and observe their performances over all topics and then over each of the topic categories separately. We also investigate the performance of five novelty detection models which use a combination of text (ASR) and each of the five high level feature combinations over all topics and each of the topic categories. Each run is compared to the performance of the baseline run. Table 6.53 and 6.54 displays the performances of the novelty detection models over all topics in both Collection_1 and Collection_2 respectively.

All high level feature runs (apart from the "People" and "ASR_People" runs), appear to perform similar to the baseline performance figures for all topics over both collections as seen in Tables 6.53 and 6.54, which means they do not aid in the detection of novel shots from within a list of shots. We observe that the "People" and "ASR_People" runs are performing below the baseline novelty performance figures over each collection, suggesting that this combination of high level features are not suitable for the detection of novel shots.

6.6.1 "General Object" Topic Category

Tables 6.55 and 6.56 display both the optimal and unbiased F-measure values of the novelty run using high level features over all topics in the "General Object" Topic category from Collection_1 and Collection_2 respectively.

It can be seen from Table 6.55 that all runs (apart from the run utilising high level features specifically combined for the Sports category, "Sports" and its combination with ASR, "ASR_Sports"), achieve performance figures similar to the baseline performance figures over the "General Object" category within Collection_1. The "Sports" run achieves an Fscore of 0.934 with a corresponding precision value of 0.882, an improvement of 0.9% and 2.0% on the baseline performance figures respectively. The combination of ASR with this high level feature combination has no effect on the novelty performance on the "General Object" category within Collection_1. We observe that the run which utilises high level features combined specifically for the "General Object" category, "General", achieves a novelty performance similar to the baseline figures. If we now look at these runs over the "General Object" category within Collection_ 2 (Table 6.56), we can see that once again the "General" run is performing similar to the baseline performance and decreases in performance during its combination with ASR. The run which utilises high level features that are combined to accurately detect sports, "Sports" once again achieves the highest novelty performance with an Fscore of 0.882, an improvement of 0.7% over the baseline Fscore figure, while the corresponding precision value outperforms the baseline by 1.4%. The combination of the run with ASR once again shows a degrading affect in the performance of the novelty detection model that uses only high level features, although this run performs above the baseline performance. We conclude that the combination of high level features specifically combined for the "Sports" category aids in the detection of novel shots from within the "General Object" category and we can clearly see that the combination of ASR with all other high level runs decrease the novelty performances.

6.6.2 "Other" Topic Category

Tables 6.57 and 6.58 display both the optimal and unbiased F-measure values of the novelty run using high level features over all topics in the "Other" category from Collection_1 and Collection_2 respectively.

Looking at Table 6.57 we observe that the highest performing runs for the "Other" topic category in Collection_1 are "ASR_General" and "ASR_Specific", runs which combine ASR with high level features specifically combined for the "General Object " category and the "Specific" category respectively. Both runs achieved an Fscore of 0.915, an increase of 0.4% on the baseline. We can clearly see that all other runs (apart from "People" and "ASR_People"), including "Other" and "ASR_Other" which combine high level features specifically for this category, achieved novelty detection performances similar to the baseline figures. The "People" and "ASR_People" runs perform below the baseline. Collection_2, Table 6.58, displays a similar trend for all runs over the "Other" category. The "People" run once again performs lower than the baseline performance suggesting that this run is not suitable for detecting novel shots for topics contained within the "Other" category in general. Once again all of the other runs which utilise high level features solely including "Other" perform similar to the baseline performance suggesting that these high level feature combinations do not perform well at detecting novel shots within the "Other" topic category. We also observe, the decrease in novelty performance during the combination of ASR with all high level feature combinations.

6.6.3 "People" Topic Category

Tables 6.59 and 6.60 display both the optimal and unbiased F-measure values of the novelty run using high level features over all topics in the "People" category from Collection_1 and Collection_2 respectively.

We observe that the "People" run which utilises high level features combined specifically for this category, is the worst performing run over both collections.

The highest performing run the "Sports" run, achieves an Fscore of 0.828 and a corresponding precision value of 0.733 over Collection_1. This is an increase of 0.6% and 1.1% over the baseline performances figures respectively. We observe that the combination of ASR with the "Sports" run, "ASR_Sports", has no effect on the novelty performance. Combining ASR with the "General" and "Specific" runs, improves each runs novelty performance to achieve an Fscore similar to that obtained by the "Sports" run, however both precision and recall values are increased. From Table 6.60 we observe that once again the "Sports" run achieves the highest Fscore of 0.727 which is an improvement of 0.3% in the performance over the baseline while also achieving an improvement of 0.3% on the precision baseline figure. All other non-ASR combined runs (apart from the "People" run) achieve the baseline performance figures. We observe that combining ASR with each of the individual runs degrades the novelty performances of each run on the "People" category over Collection_2. As a result we conclude that the novelty model that utilises high level features combined specifically for the "Sports" category achieves the highest and most consistent novelty performances over the "People" category. We also conclude that those high level features combined specifically for the "People" category, do not aid in the detection of novel shots for this topic category.

6.6.4 "Specific Object" Topic category

Tables 6.61 and 6.62 display both the optimal and unbiased F-measure values of the novelty run using features for all topics in the "Specific Object" topic category from Collection_1 and Collection_2 respectively.

From Table 6.61 we observe that once again the "Sports" run achieved the highest novelty performance achieving an Fscore of 1 and corresponding precision value of 1. This is an improvement of 12.5% and 25% on the baseline performance figures respectively. We can see that the combination of ASR with this high level feature run has no effect in the overall novelty performance. We notice that all other runs achieve a novelty performance similar to the baseline

performance. Table 6.62 displays a similar trend for each of the runs over Collection_2. Once again the "Sports" run achieves the highest Fscore of 1. As this run is consistent over both collections we conclude that it is useful in detecting novel shots in the "Specific Object" category.

6.6.5 "Sports" Topic Category

Tables 6.63 and 6.64 display both the optimal and unbiased F-measure values of the novelty run using high level features for all topics in the "Sports" topic category from Collection_1 and Collection_2 respectively.

From Table 6.63 we observe that each of the non-ASR combination runs, apart from the "People" run which performs below the baseline, achieves Fscores similar to the baseline novelty performance within Collection_1 including the "Sports" run which utilised high level features combined specifically for the "Sports" category. Combining ASR with each of the runs, improves the performance of each novelty model. The combination of ASR and features combined for each of the "General Object" category, "ASR_General" and "Specific" category, "ASR_Specific" achieve Fscores of 0.773 an increase of 0.7% on the baseline figure, while the corresponding precision figure achieves an increase of 0.9% on the baseline results. Combining ASR with the "Other" and "Sports" runs also provides an increase in the performance of novelty detection over the "Sports" category. From Table 6.64 we see that the highest performing run, which utilises high level features combined specifically for the "Other" category, "Other", achieves an Fscore of 0.720 an increase of 0.3% on the baseline performance. The runs utilising high level features combined for the "General Object" and "Sports" categories also achieve Fscores higher than the baseline figure of 0.718. Once again we see the combination of these runs with ASR degrading the performance of each novelty run. We conclude that three runs perform consistently over both collections of the "Sports" category, including those runs which utilise the high level features specifically for the "General Object", "Other" and "Sports" topic categories namely, "General", "Other" and

"Sports" respectively.

6.6.6 Summary analysis for high level features

We have carried out an extensive analysis of the performance of novelty detection models when using high level feature evidences from within video, in the identification of novel shots from within a chronologically ordered list of relevant shots firstly, over all topics and then over each of the individual topic categories separately over both collections from Tables 6.53 to 6.64 inclusively. The findings highlighted a few interesting observations including the fact that runs which utilise high level features specifically combined for the detection of shots for each of the specific topic categories do not necessarily aid in the detection of novel shots from within a list of shots relevant to a topic within that category. We observe that the detection of novel shots using high level features appears to fail when we look at each run over all topics, with high level features combined specifically for the "People" category performing worse than the baseline results detection for all topics. However as seen from the findings above the high level features combined for the "Sports" category, "Sports" run appears to consistently aid in the detection of novel shots from three of the five topic categories including "General Object", "People" and "Specific Object", while having little or no effect on the detection of novel shots for topics of the "Other" and "Sports" categories. We note that the "People" run performs badly in detecting novel shots over all topic categories.

6.7 Overall Analysis

Experiments were carried out on novelty detection models using thirty six different resource variations for each of the possible threshold values, though only a subset of the most important were included here. In this section we will provide answers for each of the questions posed at the beginning of the Chapter by presenting the trends and patterns from these results.

1. *Can novel shots be automatically detected from within a list of shots within the video domain ?*

   As illustrated from our analysis of the novelty detection runs for each of the features, it is clear there are a number of runs outperforming the baseline run which returns all shots within a list, as novel shots. This suggests that the automatic detection of novel shots from within a list of shots, within the video domain, is indeed possible. We note however, but not surprisingly, that manually disambiguated runs outperform the automatic runs.

2. *Do models designed to detect novel shots from a chronologically ordered list of shots using text resources alone outperform other resources and combinations of resources also available within the video domain or, does novelty detection need to utilise the other resources available from within video to accurately complete the task ?*

   From the analysis of the novelty models using ASR in section 6.3.6, and also from the analysis of each of the feature runs when combined with ASR, we observe that ASR is not a good feature for detecting novel shots from within a list of shots. In section 6.3.6, we observe that ASR is inconsistent over all topics and in many cases returns all shots as novel or performs worse than the baseline. When we combined ASR with other resources we observed that in many cases it reduced the performance of the original resource run. As a result we suggest that ASR should not be solely considered in determining the novelty value of a shot within a

topic. We conclude that the detection of novel shots requires the use of other resources available from within the video.

3. *How do novelty detection models developed for the identification of novel shots from a chronologically ordered list of relevant shots for a topic within the video domain, perform compared to a human assessor's performance of the task ?*

   It is desirable to designed a fully automatic novelty detection model which is able to closely match the performance of a human performing the task. As the findings illustrate, the highest novelty performance of each of the models, lie between the baseline and the human run novelty performance (apart from the novelty performance of models over the "Specific Object" topic category where the models perform similar to human performance). If we consider the novelty performances of the low level features over Collection_1 on the "General Object" topics we see that it achieves an Fscore of 0.975 with a corresponding precision value of 0.953. The minimum assessor for that topic category achieves an Fscore of 0.976 and a corresponding precision measure of 0.958. If we observe this run over Collection_2 we note however that the model is not performing as close to human performance achieving an Fscore of 0.898 when compared to the human performance of 0.940. We observe that the greatest difference between human performance and automatic novelty performances occur within the "Sports" and "People" categories suggesting that these categories are particularly difficult during the detection of novel shots.

4. *How do the performances of the many modalities available for each video sequence compare to each other in the task of detecting novel shots from a chronologically ordered list of relevant shots for a topic ?* This question will be answered in the following sections. First we will outline the best performing novelty runs over all topics and then we will look at the best performing runs for each of the five different topics categories.

6.7.1 All Topics

High level features appear to offer little or no help in the detection of novel shots when considered over all topics together. We note that each of the high level feature combinations performs similar to the baseline novelty performance suggesting that they return all shots as novel. This is consistent over both collections. We also noted that high level features combined for the "People" category appear to harm the performance of the novelty models over all topics.

It was observed that the highest performing run of the ASR resources, perform inconsistently over both collections. It achieved an Fscore similar to the baseline over Collection_1 while it performed lower than the baseline in Collection_2.

The highest performing novelty run using manual concepts over all topics and over both collections was the combination of ASR with concepts "ASR_Concepts", which achieved an Fscore of 0.872 over Collection_1 and performed similar to the baseline performance figures over Collection_2.

Low level features appear to perform well during the detection of novel shots from a list of shots. It was observed that of all the low level feature runs, two of the highest performing novelty runs included colour structure, "ColourStruc", and the combination of colour structure and edge histograms "ColourStruc_EdgeHist". "ColourStruc" achieved an Fscore of 0.893 with a corresponding precision value of 0.86, an improvement of 2.4% and 8.7% on the baseline figures respectively over Collection_1 while it achieved an Fscore of 0.822 and improvement of 1.7% on the baseline figures and precision of 0.74 over Collection_2. "ColourStruc_EdgeHist" also achieved an Fscore of 0.893 over Collection_1, however the precision is slightly less at 0.84 (an improvement of 6.3%) while it achieved an Fscore of 0.822, an improvement of 1.7% on the baseline figures and precision of 0.74 over Collection_2 (same result as ColourStruc).

Of all the features available for the detection of novel shots within a list of shots within video, it would appear from the analysis over all topics that low level

features, namely colour structure and a combination of colour structure and edge histograms, outperform all other feature runs.

6.7.2 "General Object" Category

The "Sports" run is the best consistently performing novelty run over both collections when using high level features, achieving an Fscore of 0.934 an improvement of 0.9% with a corresponding precision figure of 0.882, an improvement of 2.0% over Collection_1. It achieves an Fscore of 0.882, an improvement of 0.7% and a corresponding precision value of 0.798, an improvement of 1.4% on the baseline figures over Collection_2.

The consistently highest performing run using the low level features over the "General Object" category was the combination of colour structure and edge histograms "ColourStruc_EdgeHist" achieving an Fscore of 0.975, an increase of 5.2% on the baseline Fscore, while a corresponding precision value of 0.953 provides an improvement of 10.1% over the baseline precision figures over Collection_1. Within Collection_2, we see the "ColourStruc_EdgeHist" run achieving an Fscore of 0.898, an increase of 2.5% over the baseline Fscore, while a corresponding precision value of 0.835, provides an increase of 6% on the baseline precision figures.

We made the observation that ASR resources perform below the baseline for both Collection_1 and Collection_2 suggesting it is not a good resources for aiding in the identification of novels shots within the "General Object" category. The highest performing ASR run achieved an Fscore of 0.913 with a corresponding precision value of 0.898 over Collection_1, while within Collection_2 this run achieved an Fscore of 0.871 with a corresponding precision of 0.785.

The highest performing run for manual concepts resource over the "General Object" category was "ASR_Concepts", performing consistently well over both collections. Within Collection_1, this run achieved an Fscore of 0.946 an improvement of 2.2%, with a corresponding precision of 0.915, an improvement

of 5.8%. This run achieved an Fscore of 0.904, an improvement of 3.2% and a precision value of 0.867, an improvement of 10.2 % on the baseline performance figures over Collection_2.

We conclude that two feature resources perform well over the "General Object" category in general, including low level features in the form of colour structure and edge histograms and a combination of manually annotated concepts and ASR transcripts.

6.7.3 "Other" Category

We made the observation that a combination of ASR with each of the high level features combined specifically for the "General Object" and the "Specific Object" categories performed well with both runs achieving an Fscore of 0.915 (an improvement of 0.4%) and precision of 0.853 (an improvement of 1%) performed well for high level feature resources during novelty detection over the "Other" category within Collection_1. Within Collection_2 we have seen that all high level feature combination runs apart from the "People" run, performed similar to returning all shots as novel for the "Other" category.

The highest run of the ASR resources over Collection_1 achieved an Fscore of 0.915, an improvement of 0.4% with a corresponding precision of 0.854, an improvement of 2% on the baseline performance figures, while the same run over Collection_2 achieved an Fscore of 0.836 with a corresponding precision of 0.753. This run is performing below the baseline.

We have seen that within Collection_1 the highest performing manual concept run over the "Other" category was the combination of manual concepts and ASR resources. This run achieved an Fscore of 0.918 and a precision of 0.857, a 0.8% and 1.5% improvement on the baseline respectively, while manual concepts on their own achieve an Fscore similar to the baseline, although the precision value is increased from the baseline of 0.844 to 0.849. We have also seen within Collection_2 that manual concepts on their own, achieve the highest novelty

measure however these performance figures are below the baseline performance figures for Collection_2.

Over the low level features we observe that the combination of colour structure and edge histograms features perform consistently well over both collections. Within Collection_1, the combination of colour structure and edge histograms improved upon the baseline Fscore by 1.5% while precision improved upon the baseline by 2.4%. Within Collection_2 the run achieves an improvement of 1.3% and 2.7% upon the Fscore and precision values respectively.

We conclude that over all feature resources available, the low level colour structure and edge histograms combination should be used in the detection of novel shots from topics in the "Other" category. We suggest using a combination of colour and edge low level feature evidences, as collections can differ greatly and it is more accurate than assuming either edge or colour would be most appropriate over a certain collection.

6.7.4 "People" Category

Within the high level feature resources, we observe that the high level feature combination designed specifically for the "Sports" category consistently perform well over both collections, achieving an Fscore of 0.828 and precision of 0.732 an increase of 0.6% and 1.1% on the baseline for Collection_1, while over Collection_2 the run achieves an Fscore of 0.727 and a precision of 0.600, an improvement of 0.3% on both baseline figures respectively.

ASR resources once again performs inconsistently over both collections. We observe that the highest performing run achieves an Fscore of 0.828 an increase of 0.6% on the baseline over Collection_1, while the run performs below the baseline by 0.3% in Collection_2.

If we look at how the manual concepts aid in the detection of novelty within the "People" category, we observe that the highest performing run over both

collections is the run which uses a combination of ASR and concepts. The run achieves an Fscore of 0.831 and a precision value of 0.738, an improvement of 1% and 1.8% on the baseline performance over Collection_1, while over Collection_2 the run achieved an Fscore of 0.730 and a precision value of 0.607, an increase of 0.7% and 1.5% on the baseline figures.

We observe that low level features once again perform well over both collections within the "People" category. The run which combines colour structure and edge histograms consistently outperforms all other low level runs over both Collection_1 and Collection_2 achieving an Fscore of 0.873 and a precision value of 0.800 an increase of 6% and 10.3% on the baseline performance figures over Collection_1 while it achieved an Fscore of 0.776 and 0.695 an improvement of 7.0% and 16.22% over the baseline performance over Collection_2.

We conclude that the best use of resources for the detection of novel shots within the "People " category is the combination of two low level feature evidences, colour structure and edge histograms.

6.7.5 "Specific Object" Category

This topic category contains only one topic and as a result, it is not a good indicator of all topics that may occur in the "Specific Object" category, however we will note the results we found over this topic.

We observe that the combination of high level feature resources specifically for the "Sport" category outperforms all other high level feature runs over both collections for the "Specific Object" category achieving an Fscore of 1 and a precision value of 1 on both collection, an increase of 12.5% and 25% respectively on the baseline performances over both collections.

The highest performing ASR resources runs, perform similar to the baseline performance which returns all shots as novel for both collections by achieving an Fscore of 0.889 and precision value of 0.8.

We observe that manually annotated concepts perform consistently well over both collections within the "Specific Object" category achieving an Fscore of 1 and a precision value of 1.

We observe that both edge histogram and texture low level features perform consistently well over both collections achieving an Fscore of 1 and precision value of 1. These high performance figures were also achieved by runs using a combination of low level features including HSV colour, Canny edge and a combination of HSV colour and texture.

As illustrated from the findings presented above, there are a number of features that appear to perform well during the detection of novel shots from within the "Specific Object" category, however as there is only one topic, we cannot make a general assumption that one resources will outperform all other resources in general.

6.7.6 "Sports" Category

It was noted that high level features combined specifically for the three different topic categories, including "General object", "Other" and "Sports" perform consistently over both collections, achieving a similar or slightly higher performance than the baseline figures. Over Collection_1 all three runs achieve a novelty performance equivalent to the baseline performance of 0.768, however over Collection_2 the high level features combined specifically for the "Other" category achieves the highest Fscore of 0.720 of all runs, an improvement of 0.3%. The combination of high level features for "Sports" and "General" categories achieve an Fscore of 0.719 over the "Sports" category within Collection_2.

If we look at the performance of ASR over the "Sports" category we observe that although ASR improves upon the baseline performance within Collection_1, achieving an Fscore of 0.773 (an improvement of 0.7%), the performance of ASR resources over Collection_2 is below the baseline performance. This is consistent with all other topic categories.

Manual concepts do not perform consistently over Collection_1 and Collection_2. In Collection_1 the highest performing run, the combination of ASR and concepts achieves an Fscore of 0.785, an improvement of 2.2%. While this run, the highest performing run within Collection_2, performs below the baseline performance achieving an Fscore of 0.706.

If we look at the performance of low level features over the "Sports" category, we observe that the run which uses a combination of ASR, colour structure and edge histograms outperforms all other runs over both Collection_1 and Collection_2 achieving an Fscore of 0.839 over Collection_1 and improvement of 10.2% on the baseline performance, while it achieves an Fscore of 0.740 on Collection_2, an improvement of 3.1% on the baseline figures.

We conclude that low level features are the best resources to use during the detection of novel shots within the "Sports" category and in particular the combination of colour structure and edge histograms feature evidences.

Table 6.65 gives a summary overview of performances of each of the video resources over each of the topic categories for the detection of novel shots. Each of the runs which performs significantly better than all other runs are highlighted in bold. We can clearly see that low-level features, using the combination of colour structure and edge histograms performs well across all topic categories. The combination of ASR with each of the low-level and high-level feature resources, are not presented in this table as it has been clearly seen through the results presented above that they do not produce any measurable benefits to novelty detection within the video domain.

6.7.7 Median Difference Analysis

In this section we look at the performance of the best performing runs over each of the feature resources, and how they perform on each of the individual topics within the specific topic category against the median performance for

that topic[1].

If we consider Figure 6.1 we observe the median difference graphs for each of the best performing runs for each of the video resource features, including high level features low level features, ASR and manual concepts over the "General Object" category described earlier. Each run performs higher than the median for all topics. We observe that both manual concepts and low level features perform well over each of the topics in this category.

[1]The interested reader is directed to the Appendix for median difference graphs for each of novelty detection feature runs.

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_General	0.2 & -14.5	0.81	0.98	0.872	0.0 & -10.0	0.79	1	0.872
	0.2 & -10.0	0.81	0.98	0.872	0.2 & -10.0	0.81	0.98	0.872
	0.6 & -10.0	0.81	0.98	0.87	0.6 & -10.0	0.81	0.98	0.87
ASR_Other	0.0 & -0.5	0.79	1	0.872	0.0 & -0.5	0.79	1	0.872
	1.0 & -0.5	0.81	0.96	0.863	1.0 & -0.5	0.81	0.96	0.863
	2.0 & -0.5	0.81	0.93	0.849	2.0 & -0.5	0.81	0.93	0.849
ASR_People	0.0 & -10.5	0.78	0.92	0.821	0.0 & -19.5	0.93	0.16	0.26
	1.0 & -10.5	0.8	0.88	0.813	1.0 & -18.0	0.93	0.16	0.26
	2.0 & -10.5	0.8	0.85	0.797	1.0 & -18.5	0.93	0.16	0.26
ASR_Specific	0.2 & -6.0	0.81	0.98	0.872	0.0 & -4.0	0.79	1	0.872
	0.2 & -4.5	0.81	0.98	0.872	0.2 & -4.0	0.81	0.98	0.872
	0.6 & -5.5	0.81	0.98	0.87	0.6 & -4.0	0.81	0.98	0.87
ASR_Sports	0.0 & -17.5	0.79	1	0.872	0.0 & -17.5	0.79	1	0.872
	0.0 & -16.5	0.79	0.98	0.869	0.0 & -16.5	0.79	0.98	0.869
	0.0 & -16.0	0.79	0.98	0.867	0.0 & -17.0	0.79	0.99	0.872
General	-14.5	0.79	1	0.872	-14.5	0.79	1	0.872
	-7.5	0.78	0.93	0.829	-7.5	0.78	0.93	0.829
	-5.5	0.76	0.9	0.8	-5.5	0.76	0.9	0.8
Other	-31	0.79	1	0.872	-31	0.79	1	0.872
	-26.5	0.79	0.94	0.839	-26.5	0.79	0.94	0.839
	-20.5	0.77	0.9	0.803	-21.5	0.78	0.93	0.829
People	-20.5	0.8	0.95	0.852	-20.5	0.8	0.95	0.852
	-18	0.8	0.94	0.846	-18	0.8	0.94	0.846
	-16.5	0.78	0.92	0.821	-18.5	0.8	0.94	0.845
Specific	-6.5	0.79	1	0.872	-6.5	0.79	1	0.872
	-4.5	0.78	0.93	0.813	-4.5	0.78	0.93	0.813
	-3.5	0.79	0.87	0.787	-3.5	0.79	0.87	0.787
Sports	-30.5	0.79	1	0.872	-30.5	0.79	1	0.872
	-17.5	0.79	0.98	0.869	-17.5	0.79	0.98	0.869
	-16	0.79	0.98	0.867	-18	0.79	1	0.872

Table 6.53: Results of the Novelty detection model using high level for *all topics* over Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_General	0.0 & -10.0	0.71	0.98	0.8	0.2 & -14.5	0.72	0.96	0.796
	0.2 & -10.0	0.72	0.96	0.796	0.2 & -10.0	0.72	0.96	0.796
	0.6 & -10.0	0.72	0.95	0.793	0.6 & -10.0	0.72	0.95	0.793
ASR_Other	0.0 & -0.5	0.71	0.98	0.8	0.0 & -0.5	0.71	0.98	0.8
	1.0 & -0.5	0.72	0.94	0.789	1.0 & -0.5	0.72	0.94	0.789
	2.0 & -0.5	0.72	0.91	0.779	2.0 & -0.5	0.72	0.91	0.779
ASR_People	0.0 & -19.5	0.72	0.9	0.771	0.0 & -10.5	0.7	0.88	0.739
	1.0 & -18.5	0.72	0.89	0.767	1.0 & -10.5	0.71	0.86	0.735
	1.0 & -18.0	0.72	0.88	0.76	2.0 & -10.5	0.72	0.83	0.723
ASR_Specific	0.0 & -4.0	0.71	0.98	0.8	0.2 & -6.0	0.72	0.96	0.796
	0.2 & -4.0	0.72	0.96	0.796	0.2 & -4.5	0.72	0.96	0.796
	0.6 & -4.0	0.72	0.95	0.793	0.6 & -5.5	0.72	0.95	0.793
ASR_Sports	0.0 & -17.5	0.71	0.98	0.8	0.0 & -30.0	0.71	0.98	0.8
	0.0 & -16.5	0.71	0.97	0.798	0.0 & -16.5	0.71	0.97	0.798
	0.0 & -17.0	0.71	0.97	0.797	0.0 & -16.0	0.71	0.96	0.796
General	-14.5	0.71	1	0.808	-14.5	0.71	1	0.808
	-7.5	0.71	0.93	0.761	-7.5	0.71	0.93	0.761
	-5.5	0.7	0.9	0.739	-5.5	0.7	0.9	0.739
Other	-31	0.71	1	0.808	-31	0.71	1	0.808
	-26.5	0.71	0.94	0.771	-26.5	0.71	0.94	0.771
	-21.5	0.71	0.93	0.761	-20.5	0.7	0.9	0.741
People	-20.5	0.71	0.95	0.785	-20.5	0.71	0.95	0.785
	-18	0.71	0.93	0.777	-18	0.71	0.93	0.777
	-18.5	0.71	0.93	0.776	-16.5	0.71	0.91	0.752
Specific	-6.5	0.71	1	0.808	-6.5	0.71	1	0.808
	-4.5	0.7	0.93	0.764	-4.5	0.7	0.93	0.764
	-3.5	0.7	0.88	0.743	-3.5	0.7	0.88	0.743
Sports	-30.5	0.71	1	0.808	-30.5	0.71	1	0.808
	-18	0.71	1	0.807	-16	0.71	0.98	0.803
	-17.5	0.71	0.98	0.806	-17.5	0.71	0.98	0.806

Table 6.54: Results of the Novelty detection model using high level features for *all topics* over Collection_2

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_General	0.0 & -1.0	0.865	1.000	0.926	0.0 & -1.0	0.865	1.000	0.926
	0.2 & -1.0	0.898	0.945	0.913	1.4 & -1.0	0.897	0.930	0.905
	1.4 & -1.0	0.897	0.930	0.905	0.2 & -1.0	0.898	0.945	0.913
ASR_Other	0.0 & -0.5	0.865	1.000	0.926	0.0 & -0.5	0.865	1.000	0.926
	1.0 & -0.5	0.898	0.945	0.913	1.0 & -0.5	0.898	0.945	0.913
	2.0 & -0.5	0.897	0.888	0.874	2.0 & -0.5	0.897	0.888	0.874
ASR_People	0.0 & -10.5	0.865	1.000	0.926	1.0 & -19.5	1.000	0.282	0.430
	1.0 & -10.5	0.898	0.945	0.913	0.0 & -19.5	1.000	0.282	0.430
	0.0 & -10.0	0.875	0.953	0.907	1.0 & -10.0	0.908	0.912	0.902
ASR_Specific	0.0 & 0.0	0.865	1.000	0.926	0.0 & 0.0	0.865	1.000	0.926
	0.2 & 0.0	0.898	0.945	0.913	1.4 & 0.0	0.897	0.930	0.905
	1.4 & 0.0	0.897	0.930	0.905	0.2 & 0.0	0.898	0.945	0.913
ASR_Sports	0.0 & 6.5	0.882	1.000	0.934	0.0 & 6.5	0.882	1.000	0.934
	0.0 & 0.0	0.865	1.000	0.926	0.0 & 0.0	0.865	1.000	0.926
	1.0 & 6.5	0.915	0.945	0.922	1.0 & 6.5	0.915	0.945	0.922
General	-14.5	0.865	1.000	0.926	-14.5	0.865	1.000	0.926
					-7.5	0.865	1.000	0.926
					-5.5	0.865	1.000	0.926
Other	-31	0.865	1.000	0.926	-31.0	0.865	1.000	0.926
	-4.5	0.875	0.940	0.899	-4.5	0.875	0.940	0.899
	2.5	1.000	0.240	0.386	2.5	1.000	0.240	0.386
People	-20.5	0.865	1.000	0.926	-20.5	0.865	1.000	0.926
	-10	0.875	0.953	0.907	-10.0	0.875	0.953	0.907
	-10.5	0.875	0.940	0.899	-10.5	0.875	0.940	0.899
Specific	-6.5	0.865	1.000	0.926	-6.5	0.865	1.000	0.926
	1.5	0.958	0.515	0.664	1.5	0.958	0.515	0.664
	2	0.945	0.313	0.467	2.0	0.945	0.313	0.467
Sports	6.5	0.882	1.000	0.934	6.5	0.882	1.000	0.934
	-30.5	0.865	1.000	0.926	-30.5	0.865	1.000	0.926
	7.5	0.920	0.898	0.893	7.0	0.882	1.000	0.934

Table 6.55: Results of the Novelty detection model using high level features for the "General Object" topic category over Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_General	0 0 & -1.0	0.785	0.990	0.871	0.0 & -1.0	0.785	0.990	0.871
	1.4 & -1.0	0.822	0.932	0.863	0.2 & -1.0	0.815	0.932	0.858
	0.2 & -1.0	0.815	0.932	0.858	1.4 & -1.0	0.822	0.932	0.863
ASR_Other	0.0 & -0.5	0.785	0.990	0.871	0.0 & -0.5	0.785	0.990	0.871
	1.0 & -0.5	0.815	0.932	0.858	1.0 & -0.5	0.815	0.932	0.858
	2.0 & -0.5	0.822	0.890	0.831	2.0 & -0.5	0.822	0.890	0.831
ASR_People	1.0 & -19.5	0.822	0.932	0.863	9.0 & -2.0	0.833	0.177	0.290
	0.0 & -19.5	0.815	0.932	0.858	1.0 & -10.5	0.815	0.932	0.858
	1.0 & -10.0	0.818	0.903	0.848	1.0 & -10.5	0.815	0.932	0.858
ASR_Specific	0.0 & 0.0	0.785	0.990	0.871	0.0 & 0.0	0.785	0.990	0.871
	1.4 & 0.0	0.822	0.932	0.863	0.2 & 0.0	0.815	0.932	0.858
	0.2 & 0.0	0.815	0.932	0.858	1.4 & 0.0	0.822	0.932	0.863
ASR_Sports	0.0 & 6.5	0.797	0.990	0.878	0.0 & 6.5	0.797	0.990	0.878
	0.0 & 0.0	0.785	0.990	0.871	0.0 & 0.0	0.785	0.990	0.871
	1.0 & 6.5	0.827	0.932	0.864	1.0 & 6.5	0.827	0.932	0.864
General	-14.5	0.787	1.000	0.876	-14.5	0.787	1.000	0.876
					-7.5	0.787	1.000	0.876
					-5.5	0.787	1.000	0.876
Other	-31	0.787	1.000	0.876	-31.0	0.787	1.000	0.876
	-4.5	0.805	0.937	0.862	-4.5	0.805	0.937	0.862
	2.5	0.833	0.148	0.252	2.5	0.833	0.148	0.252
People	-20.5	0.787	1.000	0.876	-20.5	0.787	1.000	0.876
	-10	0.803	0.953	0.869	-10.0	0.803	0.953	0.869
	-10.5	0.808	0.900	0.848	-10.5	0.808	0.900	0.848
Specific	-6.5	0.787	1.000	0.876	-6.5	0.787	1.000	0.876
	1.5	0.935	0.370	0.512	1.5	0.935	0.370	0.512
	2	0.890	0.222	0.342	2.0	0.890	0.222	0.342
Sports	6.5	0.798	1.000	0.882	6.5	0.798	1.000	0.882
	-30.5	0.787	1.000	0.876	-30.5	0.787	1.000	0.876
	7	0.788	0.947	0.854	7.5	0.750	0.732	0.719

Table 6.56: Results of the Novelty detection model using high level features for the "General Object" topic category over Collection_2

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_General	0.2 & -10.0	0.853	0.993	0.915	0.2 & -10.0	0.853	0.993	0.915
	0.0 & -10.0	0.844	1.000	0.911	0.8 & -10.0	0.853	0.993	0.915
	1.2 & -10.0	0.860	0.969	0.909	0.0 & -10.0	0.844	1.000	0.911
ASR_Other	0.0 & -0.5	0.844	1.000	0.911	0.0 & -0.5	0.844	1.000	0.911
	1.0 & -0.5	0.853	0.971	0.906	1.0 & -0.5	0.853	0.971	0.906
	2.0 & -0.5	0.859	0.950	0.901	2.0 & -0.5	0.859	0.950	0.901
ASR_People	0.0 & -10.5	0.813	0.864	0.791	0.0 & -19.5	0.857	0.080	0.142
	1.0 & -10.5	0.823	0.840	0.787	1.0 & -19.5	0.857	0.080	0.142
	2.0 & -10.5	0.829	0.819	0.782	1.0 & -18.5	0.857	0.080	0.142
ASR_Specific	0.2 & -4.5	0.853	0.993	0.915	0.0 & 6.5	1.000	0.067	0.124
	0.0 & -4.0	0.844	1.000	0.911	0.0 & 6.5	1.000	0.067	0.124
	1.2 & -6.0	0.860	0.969	0.909	0.0 & -4.0	0.844	1.000	0.911
ASR_Sports	0.0 & -18.0	0.844	1.000	0.911	0.0 & -18.5	0.844	1.000	0.911
	0.0 & -17.5	0.843	0.996	0.910	0.0 & -18.0	0.844	1.000	0.911
	0.0 & -16.5	0.843	0.993	0.908	0.0 & -17.5	0.843	0.996	0.910
General	-14.5	0.844	1.000	0.911	-14.5	0.844	1.000	0.911
	-7.5	0.831	0.871	0.803	-7.5	0.831	0.871	0.803
	-5.5	0.789	0.861	0.784	-5.5	0.789	0.861	0.784
Other	-31	0.844	1.000	0.911	-31.0	0.844	1.000	0.911
	-26.5	0.836	0.894	0.833	-26.5	0.836	0.894	0.833
	-21.5	0.831	0.871	0.803	-21.5	0.831	0.871	0.803
People	-20.5	0.853	0.923	0.866	-20.5	0.853	0.923	0.866
	-18	0.851	0.920	0.862	-18.0	0.851	0.920	0.862
	-18.5	0.851	0.916	0.859	-18.5	0.851	0.916	0.859
Specific	-6.5	0.844	1.000	0.911	-6.5	0.844	1.000	0.911
	-4.5	0.806	0.876	0.805	-4.5	0.806	0.876	0.805
	-3.5	0.810	0.773	0.732	-3.5	0.810	0.773	0.732
Sports	-30.5	0.844	1.000	0.911	-30.5	0.844	1.000	0.911
	-18.5	0.843	0.996	0.910	-18.0	0.844	1.000	0.911
	-17.5	0.843	0.993	0.908	-18.5	0.843	0.996	0.910

Table 6.57: Results of the Novelty detection model using high level features for the "Other" topic category over Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_General	0.2 & -10.0	0.753	0.966	0.836	0.2 & -10.0	0.753	0.966	0.836
	0.8 & -10.0	0.753	0.963	0.835	0.0 & -10.0	0.744	0.973	0.834
	0.0 & -10.0	0.744	0.973	0.834	1.2 & -10.0	0.761	0.943	0.833
ASR_Other	0.0 & -0.5	0.744	0.973	0.834	0.0 & -0.5	0.744	0.973	0.834
	1.0 & -0.5	0.750	0.953	0.831	1.0 & -0.5	0.750	0.953	0.831
	2.0 & -0.5	0.759	0.923	0.824	2.0 & -0.5	0.759	0.923	0.824
ASR_People	0.0 & -19.5	0.753	0.883	0.779	6.0 & -10.0	0.753	0.589	0.587
	1.0 & -19.5	0.761	0.860	0.776	1.0 & -10.5	0.750	0.813	0.698
	1.0 & -18.5	0.750	0.870	0.773	2.0 & -10.5	0.759	0.784	0.692
ASR_Specific	0.2 & -4.0	0.753	0.966	0.836	0.2 & -4.5	0.753	0.966	0.836
	0.8 & -4.0	0.753	0.963	0.835	0.0 & -4.0	0.744	0.973	0.834
	0.0 & -4.0	0.744	0.973	0.834	1.2 & -6.0	0.761	0.943	0.833
ASR_Sports	0.0 & -18.5	0.744	0.973	0.834	0.0 & -18.0	0.744	0.971	0.833
	0.0 & -18.0	0.744	0.971	0.833	0.0 & -17.5	0.744	0.970	0.832
	0.0 & -17.5	0.744	0.970	0.832	0.0 & -16.5	0.746	0.966	0.831
General	-14.5	0.751	1.000	0.847	-14.5	0.751	1.000	0.847
	-7.5	0.751	0.864	0.722	-7.5	0.751	0.864	0.722
	-5.5	0.751	0.859	0.713	-5.5	0.751	0.859	0.713
Other	-31	0.751	1.000	0.847	-31.0	0.751	1.000	0.847
	-26.5	0.751	0.886	0.755	-26.5	0.751	0.886	0.755
	-21.5	0.751	0.864	0.722	-21.5	0.751	0.864	0.722
People	-20.5	0.751	0.916	0.789	-20.5	0.751	0.916	0.789
	-18	0.751	0.913	0.787	-18.0	0.751	0.913	0.787
	-18.5	0.751	0.910	0.784	-18.5	0.751	0.910	0.784
Specific	-6.5	0.751	1.000	0.847	-6.5	0.751	1.000	0.847
	-4.5	0.756	0.907	0.799	-4.5	0.756	0.907	0.799
	-3.5	0.756	0.837	0.754	-3.5	0.756	0.837	0.754
Sports	-30.5	0.751	1.000	0.847	-30.5	0.751	1.000	0.847
	-18	0.751	0.999	0.846	-18.5	0.751	0.997	0.845
	-18.5	0.751	0.997	0.845	-17.5	0.753	0.993	0.844

Table 6.58: Results of the Novelty detection model using high level features for the "Other" topic category over Collection_2

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_General	0.2 & -10.0	0.738	0.995	0.828	0.0 & -10.0	0.725	1.000	0.823
	0.0 & -10.0	0.725	1.000	0.823	0.2 & -10.0	0.738	0.995	0.828
	0.6 & -10.0	0.737	0.980	0.821	0.6 & -10.0	0.737	0.980	0.821
ASR_Other	0.0 & -0.5	0.725	1.000	0.823	0.0 & -0.5	0.725	1.000	0.823
	1.0 & -0.5	0.733	0.955	0.807	1.0 & -0.5	0.733	0.955	0.807
	2.0 & -0.5	0.738	0.923	0.796	2.0 & -0.5	0.738	0.923	0.796
ASR_People	0.0 & -10.5	0.727	0.847	0.768	0.0 & -19.5	0.917	0.083	0.150
	1.0 & -10.5	0.737	0.807	0.755	1.0 & -18.5	0.917	0.083	0.150
	2.0 & -10.5	0.742	0.775	0.738	1.0 & -19.5	0.917	0.083	0.150
ASR_Specific	0.2 & -3.0	0.738	0.995	0.828	0.0 & -3.0	0.725	1.000	0.823
	0.0 & -3.0	0.725	1.000	0.823	0.2 & -3.0	0.738	0.995	0.828
	0.6 & -3.0	0.737	0.980	0.821	0.6 & -3.0	0.737	0.980	0.821
ASR_Sports	0.0 & -17.0	0.733	0.985	0.828	0.0 & -17.5	0.727	1.000	0.824
	0.0 & -17.5	0.727	1.000	0.824	0.0 & -18.0	0.725	1.000	0.823
	0.0 & -18.0	0.725	1.000	0.823	0.0 & -16.5	0.735	0.937	0.815
General	-14.5	0.725	1.000	0.823	-14.5	0.725	1.000	0.823
	-7.5	0.712	0.888	0.784	-7.5	0.712	0.888	0.784
	-5.5	0.657	0.780	0.695	-5.5	0.657	0.780	0.695
Other	-31	0.725	1.000	0.823	-31.0	0.725	1.000	0.823
	-26.5	0.715	0.897	0.789	-26.5	0.715	0.897	0.789
	-21.5	0.712	0.888	0.784	-21.5	0.712	0.888	0.784
People	-20.5	0.732	0.905	0.802	-20.5	0.732	0.905	0.802
	-18.5	0.733	0.863	0.781	-10.0	0.775	0.708	0.736
	-16.5	0.727	0.847	0.768	-18.5	0.733	0.863	0.781
Specific	-6.5	0.725	1.000	0.823	-6.5	0.725	1.000	0.823
	-3.5	0.745	0.912	0.810	-3.5	0.745	0.912	0.810
	-2	0.737	0.920	0.809	-2.5	0.728	0.872	0.784
Sports	-17	0.733	0.985	0.828	-18.5	0.727	1.000	0.824
	-18.5	0.727	1.000	0.824	-30.5	0.725	1.000	0.823
	-30.5	0.725	1.000	0.823	-17.5	0.735	0.937	0.815

Table 6.59: Results of the Novelty detection model using high level features for the "People" topic category Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_General	0.0 & -10.0	0.598	0.997	0.723	0.2 & -10.0	0.602	0.960	0.717
	0.2 & -10.0	0.602	0.960	0.717	0.0 & -10.0	0.598	0.997	0.723
	0.6 & -10.0	0.597	0.935	0.705	0.6 & -10.0	0.597	0.935	0.705
ASR_Other	0.0 & -0.5	0.598	0.997	0.723	0.0 & -0.5	0.598	0.997	0.723
	1.0 & -0.5	0.593	0.917	0.698	1.0 & -0.5	0.593	0.917	0.698
	2.0 & -0.5	0.597	0.898	0.695	2.0 & -0.5	0.597	0.898	0.695
ASR_People	0.0 & -19.5	0.598	0.855	0.689	5.0 & -10.0	0.655	0.647	0.646
	1.0 & -18.5	0.598	0.840	0.681	1.0 & -10.5	0.583	0.768	0.646
	1.0 & -19.5	0.597	0.820	0.677	2.0 & -10.5	0.582	0.738	0.637
ASR_Specific	0.0 & -3.0	0.598	0.997	0.723	0.2 & -3.0	0.602	0.960	0.717
	0.2 & -3.0	0.602	0.960	0.717	0.0 & -3.0	0.598	0.997	0.723
	0.6 & -3.0	0.597	0.935	0.705	0.6 & -3.0	0.597	0.935	0.705
ASR_Sports	0.0 & -17.5	0.600	0.997	0.725	0.0 & -17.0	0.598	0.947	0.716
	0.0 & -18.0	0.598	0.997	0.723	0.0 & -17.5	0.600	0.997	0.725
	0.0 & -16.5	0.603	0.937	0.720	0.0 & -18.0	0.598	0.997	0.723
General	-14.5	0.598	1.000	0.725	-14.5	0.598	1.000	0.725
	-7.5	0.588	0.887	0.692	-7.5	0.588	0.887	0.692
	-5.5	0.550	0.767	0.616	-5.5	0.550	0.767	0.616
Other	-31	0.598	1.000	0.725	-31.0	0.598	1.000	0.725
	-26.5	0.587	0.892	0.693	-26.5	0.587	0.892	0.693
	-21.5	0.588	0.887	0.692	-21.5	0.588	0.887	0.692
People	-20.5	0.600	0.898	0.704	-20.5	0.600	0.898	0.704
	-10	0.640	0.760	0.691	-18.5	0.592	0.838	0.678
	-18.5	0.592	0.838	0.678	-16.5	0.585	0.827	0.669
Specific	-6.5	0.598	1.000	0.725	-6.5	0.598	1.000	0.725
	-3.5	0.595	0.882	0.698	-3.5	0.595	0.882	0.698
	-2.5	0.598	0.873	0.695	-2.0	0.592	0.865	0.690
Sports	-18.5	0.600	1.000	0.727	-17.0	0.598	0.950	0.718
	-30.5	0.598	1.000	0.725	-18.5	0.600	1.000	0.727
	-17.5	0.603	0.940	0.721	-30.5	0.598	1.000	0.725

Table 6.60: Results of the Novelty detection model using high level features for the "People" topic category over Collection_2

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T**	**Prec**	**Recall**	**Fscore**
ASR_General	0.0 & -1.0	0.800	1.000	0.889	0.0 & -1.0	0.800	1.000	0.889
	10.0 & -1.0	0.750	0.750	0.750	10.0 & -1.0	0.750	0.750	0.750
					0.2 & -1.0	0.800	1.000	0.889
ASR_Other	0.0 & -0.5	0.800	1.000	0.889	0.0 & -0.5	0.800	1.000	0.889
	5.0 & -0.5	0.750	0.750	0.750	5.0 & -0.5	0.750	0.750	0.750
	0.0 & 0.0	1.000	0.250	0.400	0.0 & 0.0	1.000	0.250	0.400
ASR_People	0.0 & -10.0	0.800	1.000	0.889	0.0 & -19.5	1.000	0.500	0.667
	5.0 & -10.0	0.750	0.750	0.750	5.0 & -10.0	0.750	0.750	0.750
	0.0 & 0.0	1.000	0.500	0.667	0.0 & -19.0	1.000	0.500	0.667
ASR_Specific	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	0.0 & 1.5	1.000	0.750	0.857	0.0 & 1.5	1.000	0.750	0.857
	10.0 & 0.0	0.750	0.750	0.750	10.0 & 0.0	0.750	0.750	0.750
ASR_Sports	0.0 & 8.0	1.000	1.000	1.000	0.0 & 8.0	1.000	1.000	1.000
	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	5.0 & 8.0	1.000	0.750	0.857	5.0 & 8.0	1.000	0.750	0.857
General	-14.5	0.800	1.000	0.889	-14.5	0.800	1.000	0.889
					-14.5	0.800	1.000	0.889
					-5.5	0.800	1.000	0.889
Other	-31	0.800	1.000	0.889	-31.0	0.800	1.000	0.889
	2.5	1.000	0.250	0.400	2.5	1.000	0.250	0.400
					-31.0	0.800	1.000	0.889
People	-20.5	0.800	1.000	0.889	-20.5	0.800	1.000	0.889
	-4.5	1.000	0.500	0.667	-4.5	1.000	0.500	0.667
	4.5	1.000	0.250	0.400	4.5	1.000	0.250	0.400
Specific	-6.5	0.800	1.000	0.889	-6.5	0.800	1.000	0.889
	1.5	1.000	0.750	0.857	1.5	1.000	0.750	0.857
	2	1.000	0.500	0.667	2.0	1.000	0.500	0.667
Sports	8	1.000	1.000	1.000	8.0	1.000	1.000	1.000
	-30.5	0.800	1.000	0.889	-30.5	0.800	1.000	0.889
	9	1.000	0.500	0.667	9.0	1.000	0.500	0.667

Table 6.61: Results of the Novelty detection model using high level for the "Specific Object" topic category features over Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_General	0.0 & -1.0	0.800	1.000	0.889	0.0 & -1.0	0.800	1.000	0.889
	10.0 & -1.0	0.750	0.750	0.750	10.0 & -1.0	0.750	0.750	0.750
					2.0 & -10.0	0.800	1.000	0.889
ASR_Other	0.0 & -0.5	0.800	1.000	0.889	0.0 & -0.5	0.800	1.000	0.889
	5.0 & -0.5	0.750	0.750	0.750	5.0 & -0.5	0.750	0.750	0.750
	0.0 & 0.0	1.000	0.250	0.400	0.0 & 0.0	1.000	0.250	0.400
ASR_People	0.0 & -19.5	0.800	1.000	0.889	9.0 & -10.0	0.750	0.750	0.750
	5.0 & -10.0	0.750	0.750	0.750	5.0 & -10.0	0.750	0.750	0.750
	0.0 & -19.0	1.000	0.500	0.667	5.0 & -10.0	0.750	0.750	0.750
ASR_Specific	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	0.0 & 1.5	1.000	0.750	0.857	0.0 & 1.5	1.000	0.750	0.857
	10.0 & 0.0	0.750	0.750	0.750	10.0 & 0.0	0.750	0.750	0.750
ASR_Sports	0.0 & 8.0	1.000	1.000	1.000	0.0 & 8.0	1.000	1.000	1.000
	0.0 & 0.0	0.800	1.000	0.889	0.0 & 0.0	0.800	1.000	0.889
	5.0 & 8.0	1.000	0.750	0.857	5.0 & 8.0	1.000	0.750	0.857
General	-14.5	0.800	1.000	0.889	-14.5	0.800	1.000	0.889
					-7.5	0.800	1.000	0.889
					-7.5	0.800	1.000	0.889
Other	-31	0.800	1.000	0.889	-31.0	0.800	1.000	0.889
	2.5	1.000	0.250	0.400	2.5	1.000	0.250	0.400
					-4.5	0.800	1.000	0.889
People	-20.5	0.800	1.000	0.889	-20.5	0.800	1.000	0.889
	-4.5	1.000	0.500	0.667	-4.5	1.000	0.500	0.667
	4.5	1.000	0.250	0.400	4.5	1.000	0.250	0.400
Specific	-6.5	0.800	1.000	0.889	-6.5	0.800	1.000	0.889
	1.5	1.000	0.750	0.857	1.5	1.000	0.750	0.857
	2	1.000	0.500	0.667	2.0	1.000	0.500	0.667
Sports	8	1.000	1.000	1.000	8.0	1.000	1.000	1.000
	-30.5	0.800	1.000	0.889	-30.5	0.800	1.000	0.889
	9	1.000	0.500	0.667	9.0	1.000	0.500	0.667

Table 6.62: Results of the Novelty detection model using high level features for the "Specific Object" topic category over Collection_2

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T**	**Prec**	**Recall**	**Fscore**
ASR_General	1.8 & -10.0	0.663	0.983	0.773	0.8 & -1.0	0.653	0.977	0.763
	1.6 & -10.0	0.663	0.983	0.772	0.8 & -10.0	0.657	0.983	0.766
	2.0 & -10.0	0.663	0.977	0.771	0.2 & -1.0	0.657	0.983	0.765
ASR_Other	2.0 & -0.5	0.663	0.977	0.771	1.0 & -0.5	0.660	0.983	0.768
	1.0 & -0.5	0.660	0.983	0.768	0.0 & -0.5	0.657	1.000	0.768
	0.0 & -0.5	0.657	1.000	0.768	2.0 & -0.5	0.663	0.977	0.771
ASR_People	2.0 & -10.5	0.660	0.970	0.768	0.0 & -19.5	1.000	0.173	0.280
	1.0 & -10.5	0.657	0.977	0.765	1.0 & -10.0	0.653	0.960	0.759
	0.0 & -10.5	0.653	0.993	0.765	1.0 & -10.5	0.657	0.977	0.765
ASR_Specific	1.8 & -6.0	0.663	0.983	0.773	0.0 & 2.0	1.000	0.087	0.155
	1.6 & -6.0	0.663	0.983	0.772	0.0 & 2.0	1.000	0.087	0.155
	2.0 & -6.0	0.663	0.977	0.771	1.0 & -4.0	0.657	0.993	0.768
ASR_Sports	2.0 & -17.0	0.663	0.977	0.771	1.0 & -16.0	0.653	0.970	0.762
	1.0 & -17.0	0.660	0.983	0.768	1.0 & 0.0	0.657	0.977	0.765
	2.0 & 0.0	0.660	0.970	0.768	1.0 & -17.0	0.660	0.983	0.768
General	-14.5	0.657	1.000	0.768	-7.5	0.653	0.993	0.765
	-7.5	0.653	0.993	0.765	-14.5	0.657	1.000	0.768
					-5.5	0.653	0.993	0.765
Other	-31	0.657	1.000	0.768	-4.5	0.650	0.977	0.759
	-26.5	0.653	0.993	0.765	-26.5	0.653	0.993	0.765
	-4.5	0.650	0.977	0.759	-31.0	0.657	1.000	0.768
People	-20.5	0.653	0.993	0.765	-10.5	0.650	0.977	0.759
	-10.5	0.650	0.977	0.759	-20.5	0.653	0.993	0.765
	-4	1.000	0.180	0.293	-4.0	1.000	0.180	0.293
Specific	-6.5	0.657	1.000	0.768	-6.5	0.657	1.000	0.768
	-4.5	0.680	0.737	0.563	-3.5	0.677	0.730	0.560
	-3.5	0.677	0.730	0.560	-4.5	0.680	0.737	0.563
Sports	-30.5	0.657	1.000	0.768	-16.0	0.653	0.983	0.762
	-17.5	0.653	0.993	0.765	-17.5	0.653	0.993	0.765
	1.5	0.677	0.970	0.764	-30.5	0.657	1.000	0.768

Table 6.63: Results of the Novelty detection model using high level features for the "Sports" topic category over Collection_1

	Optimised Results				Unbiased Results			
Runs	**T.**	**Prec**	**Recall**	**Fscore**	**T.**	**Prec**	**Recall**	**Fscore**
ASR_General	0.8 & -1.0	0.670	0.973	0.708	1.8 & -10.0	0.663	0.913	0.697
	0.8 & -10.0	0.667	0.973	0.708	1.6 & -10.0	0.663	0.913	0.696
	0.2 & -1.0	0.667	0.973	0.707	2.0 & -10.0	0.663	0.913	0.698
ASR_Other	1.0 & -0.5	0.667	0.953	0.705	2.0 & -0.5	0.663	0.913	0.698
	0.0 & -0.5	0.667	0.973	0.704	1.0 & -0.5	0.667	0.953	0.705
	2.0 & -0.5	0.663	0.913	0.698	0.0 & -0.5	0.667	0.973	0.704
ASR_People	0.0 & -19.5	0.667	0.973	0.707	2.0 & -10.5	0.663	0.913	0.699
	1.0 & -10.0	0.667	0.953	0.707	1.0 & -10.5	0.667	0.953	0.705
	1.0 & -10.5	0.667	0.953	0.705	1.0 & -10.5	0.667	0.953	0.705
ASR_Specific	0.8 & -4.0	0.667	0.973	0.708	1.8 & -6.0	0.663	0.913	0.697
	0.2 & -4.0	0.667	0.973	0.707	1.6 & -6.0	0.663	0.913	0.696
	1.0 & -4.0	0.667	0.953	0.705	2.0 & -6.0	0.663	0.913	0.698
ASR_Sports	1.0 & -16.0	0.667	0.953	0.706	2.0 & -17.0	0.663	0.913	0.698
	1.0 & 0.0	0.667	0.953	0.705	1.0 & -17.0	0.667	0.953	0.705
	1.0 & -17.0	0.667	0.953	0.705	2.0 & 0.0	0.663	0.913	0.699
General	-7.5	0.667	1.000	0.719	-14.5	0.667	1.000	0.718
	-14.5	0.667	1.000	0.718	-7.5	0.667	1.000	0.719
					-7.5	0.667	1.000	0.719
Other	-4.5	0.667	1.000	0.720	-31.0	0.667	1.000	0.718
	-26.5	0.667	1.000	0.719	-26.5	0.667	1.000	0.719
	-31	0.667	1.000	0.718	-4.5	0.667	1.000	0.720
People	-10.5	0.665	0.998	0.717	-20.5	0.663	0.998	0.716
	-20.5	0.663	0.998	0.716	-10.5	0.665	0.998	0.717
	-4	0.750	0.153	0.254	-4.0	0.750	0.153	0.254
Specific	-6.5	0.667	1.000	0.718	-6.5	0.667	1.000	0.718
	-3.5	0.600	0.717	0.492	-4.5	0.600	0.717	0.492
	-4.5	0.600	0.717	0.492	-3.5	0.600	0.717	0.492
Sports	-16	0.667	1.000	0.719	-30.5	0.667	1.000	0.718
	-17.5	0.667	1.000	0.719	-17.5	0.667	1.000	0.719
	-30.5	0.667	1.000	0.718	1.5	0.660	0.950	0.691

Table 6.64: Results of the Novelty detection model using high level features for the "Sport" topic category over Collection_2

		General Obj. (F,P)	Other (F,P)	People (F,P)	Specific Obj. (F,P)	Sports (F,P)
High-level	Col_1	+0.9%, +2.0%	+0.4%, +1%	+0.6%, +1.1%	+12.5%, +25%	0%, 0%
	Col_2	+0.7%, +1.4%	0%, 0%	+0.3%, +0.3%	+12.5%, +25%	+0.3%, 0%
Low-level	Col_1	+5.2%, +10.1%	+1.5%, +2.4%	+6%, +10.3%	+12.5, +25%	+10.2%, +17.1%
	Col_2	+2.5%, +0%	+1.3%, +2.7%	+7.0%, +16.22%	+12.5%, +25%	+3.1%, +11.4%
ASR	Col_1	−1.4%, +3.8%	+0.4%, +2%	+0.6%, +1.8%	0%, 0%	+0.7%, +0.9%
	Col_2	−0.6%, −2.3%	−1.3%, +0.3%	−0.3%, 0%	0%, 0%	−1.4%, 0%
Concepts	Col_1	−0.9%, 0%	0%, +0.6%	+0.1%, +0.3%	+12.5%, +25%	−1.0%, +1.2%
	Col_2	−4.5%, +0.4%	−0.3%, +0.4%	0%, +0.3%	+12.5%, +25%	−4.0%, 0%
ASR & Concepts	Col_1	+2.2%, +5.8%	+0.8%, +1.5%	+1%, +1.8%	0%, 0%	+2.2%, +9.1%
	Col_2	+3.2%, +10.2%	−1.1%, +0.8%	+0.7%, +1.5%	0%, 0%	−1.7%, 0%

Table 6.65: Summary of the overall effects of video resources on the detection of novel shots over each topic category. Each cell contains the percentage increase or decrease on each of the Fscore and precision baseline figures respectively.

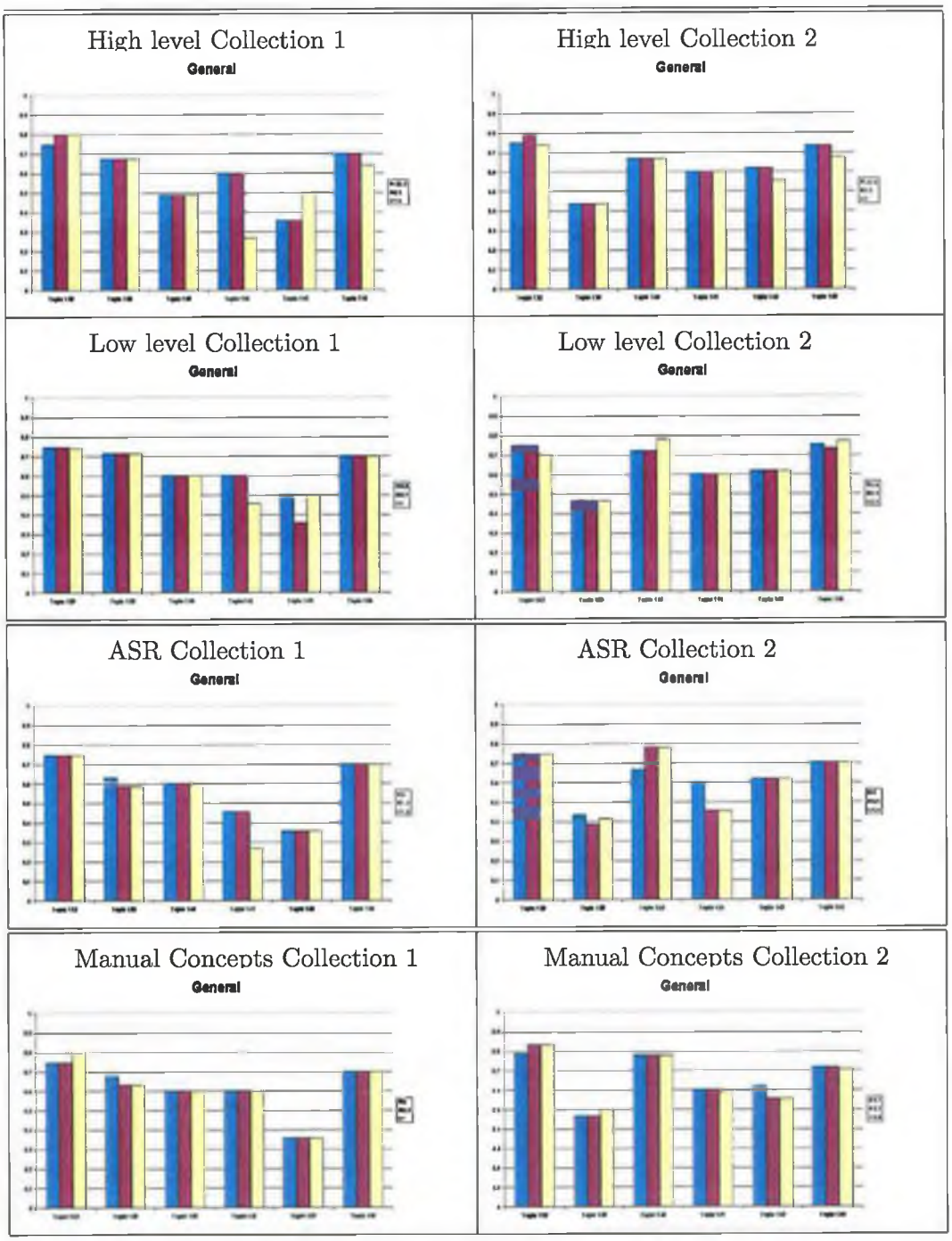

Figure 6.1: Median difference graphs over the "General Object" Category

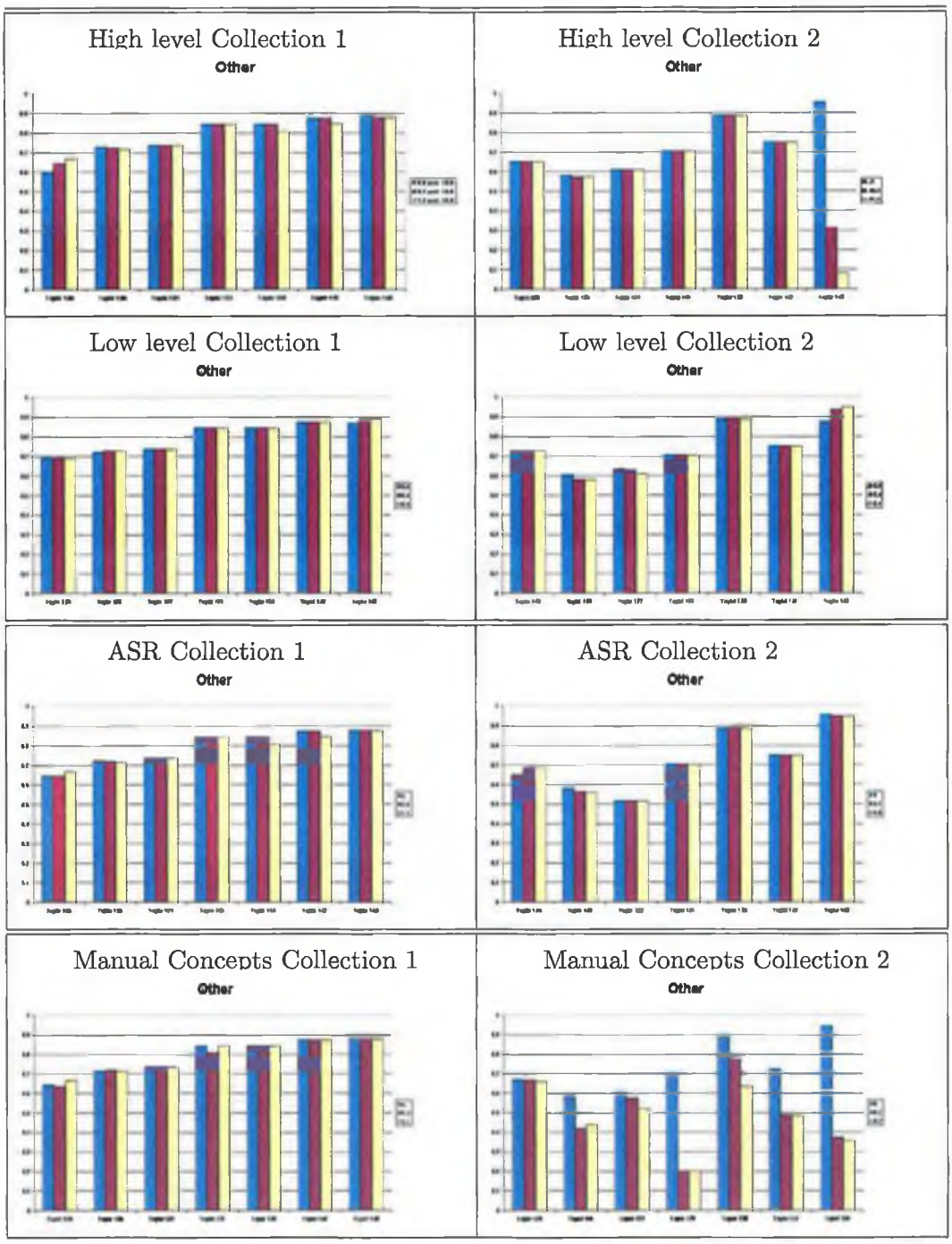

Figure 6.2: Median difference graphs over the "Other" Category

If we consider Figure 6.2, we observe each of the highest performing runs from with each resources performs higher than the median for all topics over the "Other" topic category. We observe that over Collection_1, there is very little difference in the performance of each run over each topic, however of those that vary slightly we observe that low level features perform well. This trend can be more clearly seen within Collection_2 where three of the seven topics perform better using low level resources, while each of the other topics achieve a novelty performance similar using any of the resources available.

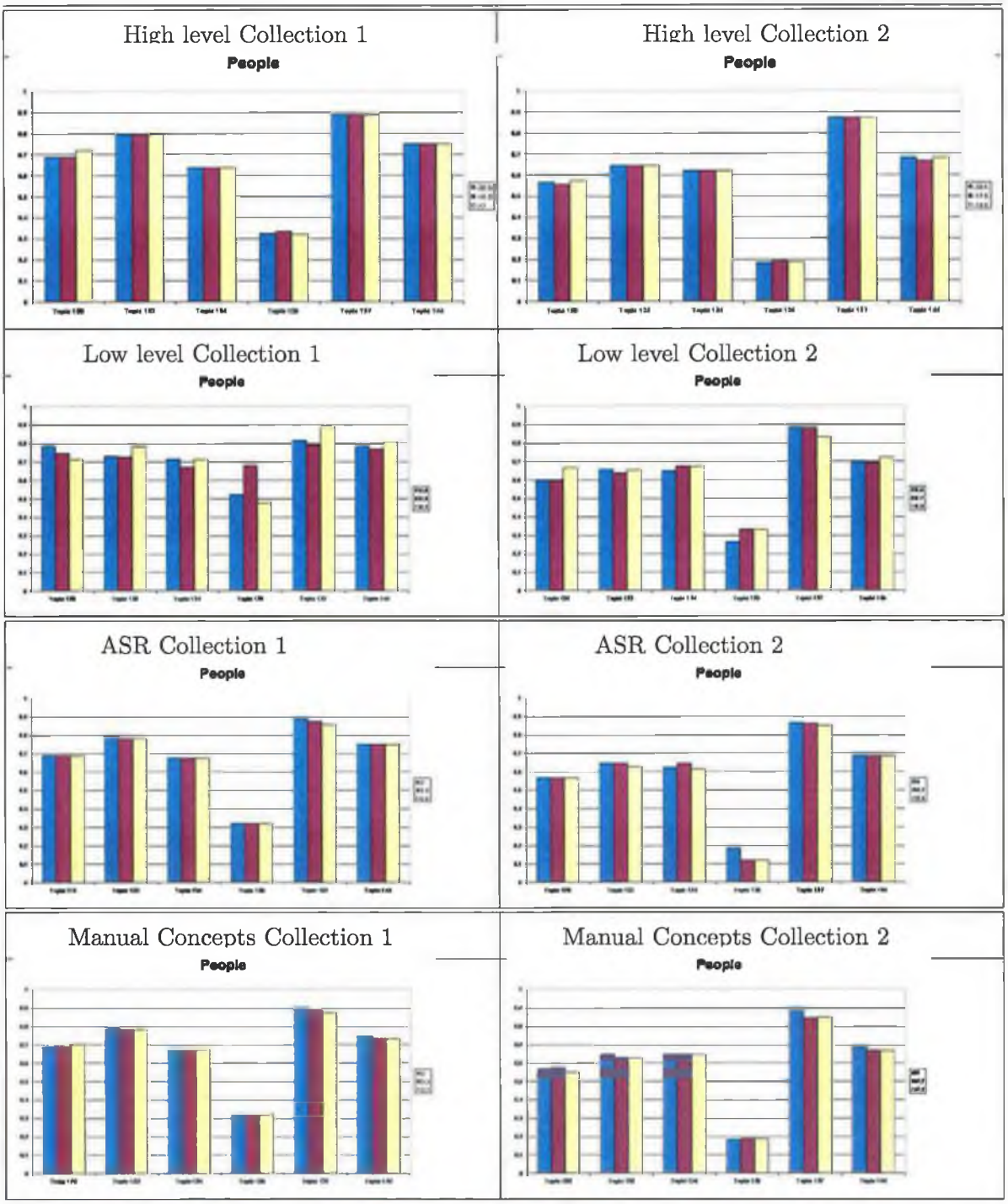

Figure 6.3: Median difference graphs over the "People" Category

If we consider Figure 6.3 once again we observe that each of the highest performing runs for novelty detection within the "People" category perform higher than the median for all topics. We observe that the highest performing low level feature run "ColourStruc_EdgeHist" consistently performs well over each of the topics (apart from Topic 133, defined as "Find shots of Saddam Hussein", where this run performed below each of the other novelty runs within Collection_1), over each of the two collections.

Figure 6.4 shows the median difference graphs for each of the different highest performing runs from each of the video resources over the "Specific Object" topic category over Collection_1 and Collection_2 respectively. As there is only one topic in the "Specific Object" category, once again we cannot generalise for all topics that may be considered to belong to this category, however we observe that each of the highest performing runs for novelty detection within the "Specific Object" category achieved a novelty performance higher than the median Fscore values. Topic 129 which is defined as "Find shots zooming in on the U.S Capitol dome" has between 70 and 90% of its shots considered novel. We observe that each of the resources (apart from ASR) achieve a high performing novel score on this topic.

Figure 6.5 shows the median difference graphs for each of the topics in the "Sports" category. Each of the novelty runs for each of the resources perform above the median. We observe however that Topic 130 defined as "Find shots of a hockey rink with at least one of the nets fully visible from some point of view." performs just above the median Fscore results. We note that low level features perform well over all topics over both collections.

If we now consider the topics which only contain up to 70% of novel shots within each topic over both collections from Table 6.3 and Table 6.4 respectively, we observe that once again low level features consistently perform highly over each of the topics. We notice however that for Topic 126 in Collection_1 and Topics 139 and 131 in Collection_2, manually annotated concepts achieves the highest novelty performance.

We observe that each of the resources perform equally well on the topics where all shots are considered novel over both collections. This characteristic holds when we consider each of the topics from both collection that are categories in to the 90-100% novel range. We note that the performance of resources on topics within the 70-90% range varying widely and this is due to the topic category to which each belong.

Threshold variations

The threshold variation graphs show the curves of the F-measure, precision and recall values as the thresholds are varied as part of the experimental run under investigation. In Chapter 4 section 3 we discussed the need for threshold values due to the varying tolerance levels of different humans to the presentation of redundant information. The graphs allow us to visually observe the effects of precision, recall and F-measure values when the threshold values vary from the extreme of allowing all non-duplicate shots to be considered novel to the other extreme of returning shots that only contain all novel information. Figure E.14 shows the threshold variation graph of the highest performing low level feature run "ColourStruc_EdgeHist" over each of the topic categories. We can clearly see that as the threshold increases, the F-measure and recall curves decrease while the precision value curve increases. This illustrates the importance of the threshold values. Users wishing to see only novel shots will be interested in choosing high threshold values, while users wishing to receive as many novel shots as possible but at the same time return as many shots as possible will be interested in threshold values where the F-measure curve peaks. Each of the threshold variation graphs for each of the runs are provided in the Appendix for the interested reader.

6.8 Summary

At the beginning of this Chapter we set out to answer four research questions to the new problem of the detection of novel shots from within a chronologically ordered list of known relevant shots to a topic, in the video domain. This was achieved by investigating the performance of the individual utilisation of each of the four different resources associated with video, namely, text, low-level feature evidences, high-level feature evidences and manually annotated concept descriptions, and their various combinations in novelty detection models. Each section displays Tables that presented both the optimal and unbiased F-measure values achieved by each run, over all topics as a whole and on each topic category separately, across both Collection_1 and Collection_2.

We have seen that low level features perform best in our experiments on both collections. As video is so diverse in colour, shapes and motion it is therefore necessary to use both colour structure and edge histogram resources available to achieve the best overall performances of novelty detection.

We have observed that ASR transcripts do not aid in the detection of novel shots within a list of chronologically ordered shots for a specific topic within the video domain.

We observed that manual concepts can aid the detection of novel shots over some topics when combined with ASR transcripts. This combination is necessary as many shots are labeled with over used concepts, due to the limited number of concepts in the ontology to describe the contents of the shot and as a result many shots which may visually appear different are considered redundant. The addition of ASR reduces this redundancy.

High level features also performed above the baseline over many topics although this was not evident during the analysis over all topics together. However they do not perform as well as low level features or the combination of ASR and manually annotated concepts.

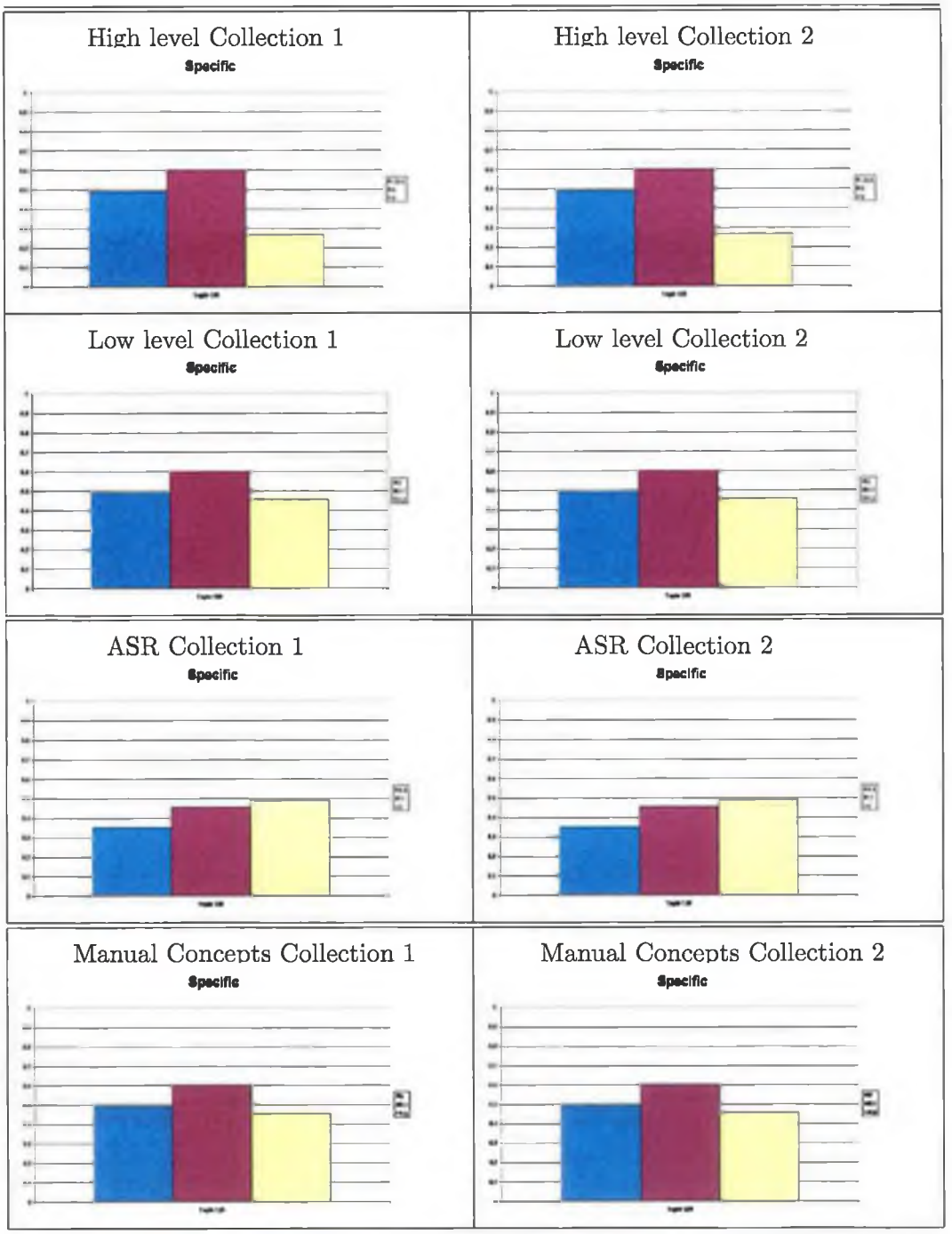

Figure 6.4: Median difference graphs over the "Specific Object" Category

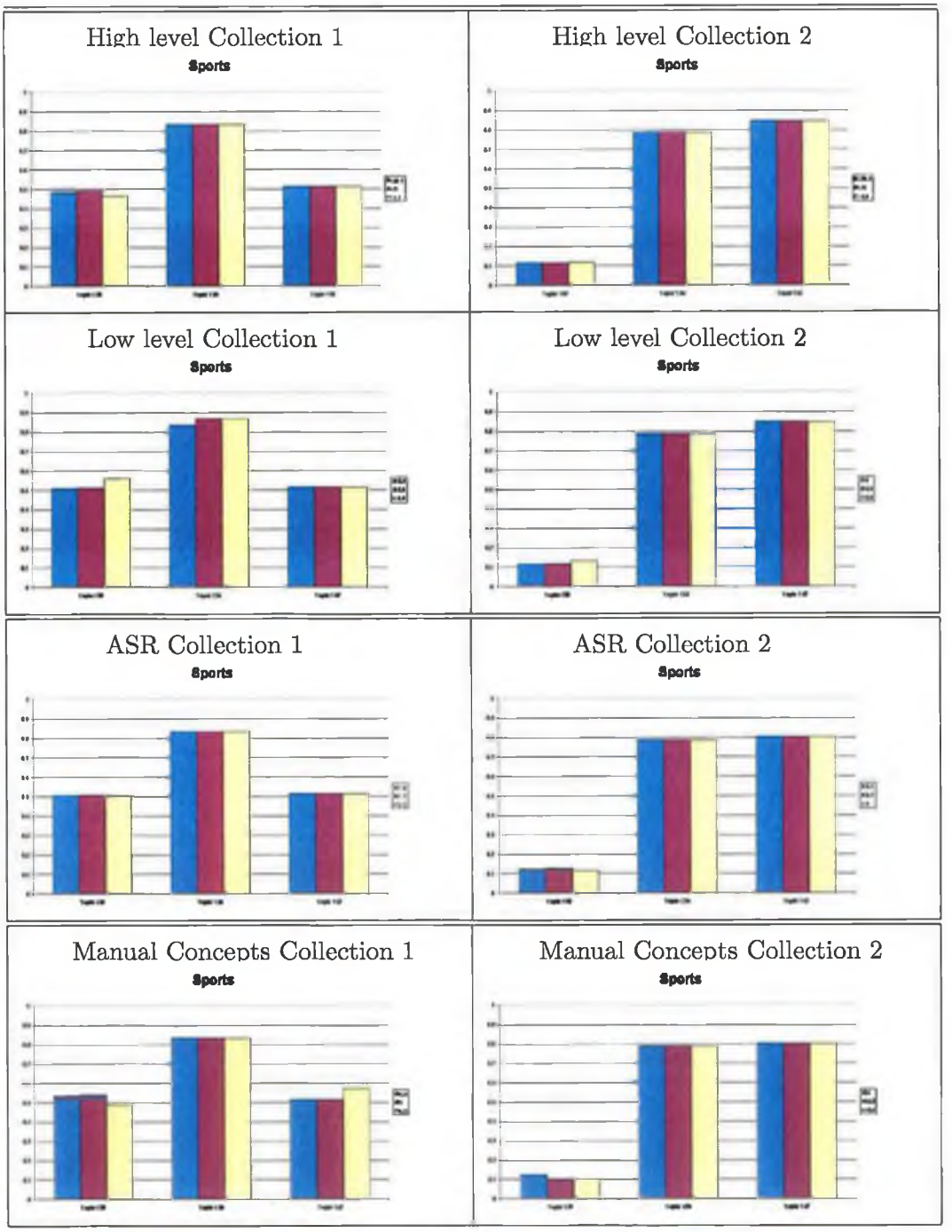

Figure 6.5: Median difference graphs over the "Sports" Category

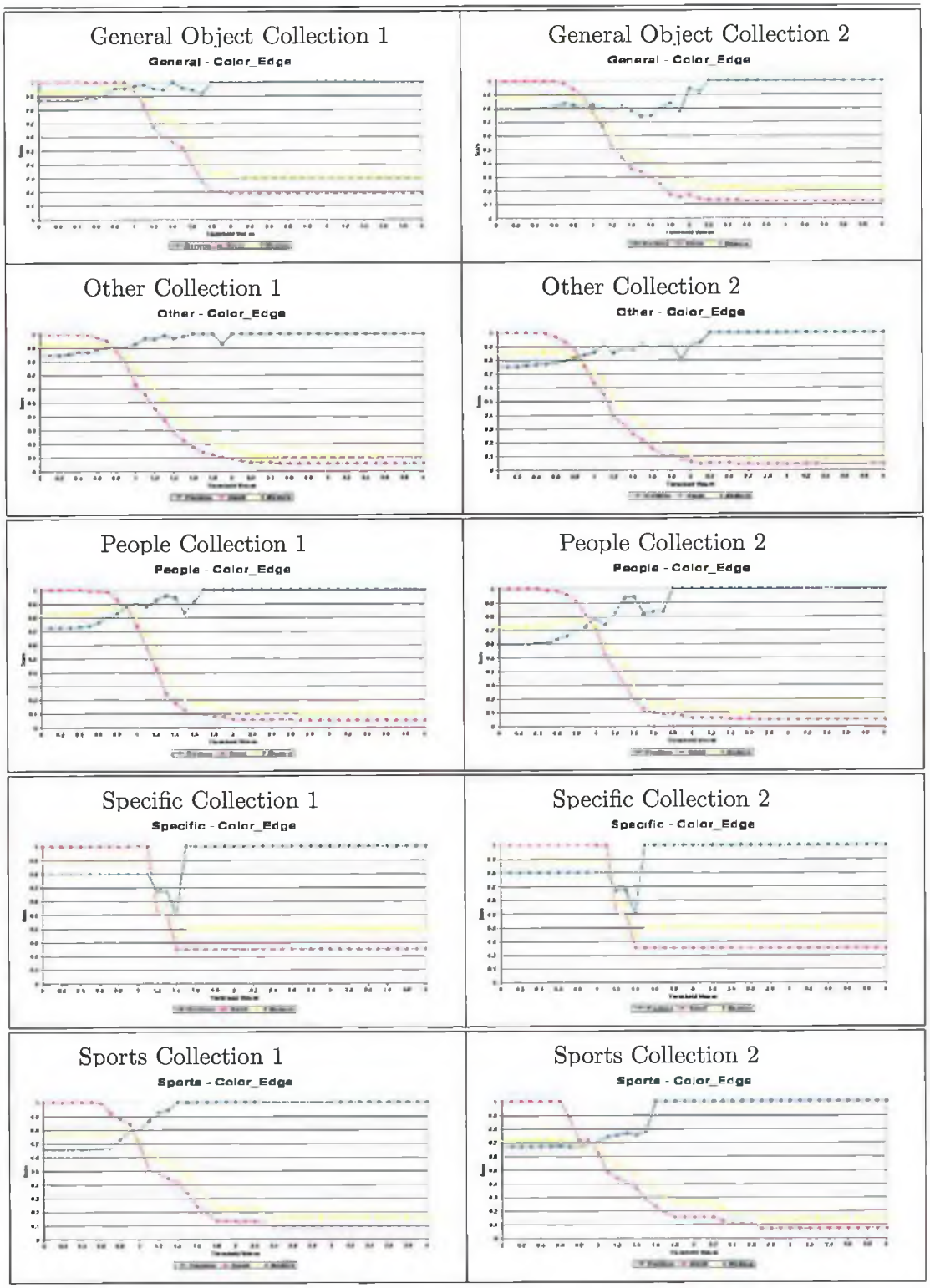

Figure 6.6: Threshold variation graphs over the both Collection_1 and Collection_2

Chapter 7

Conclusions

This chapter will briefly review Chapters 1 to 6 before presenting our general conclusions. It will then outline some ideas for possible future work and finally conclude with a brief final summary.

7.1 Summary of Thesis

In this thesis we presented the concept of, and evaluated the effectiveness of, models designed to detect novel shots from within a chronological list of known relevant shots for a particular user information need.

The work presented is a new concept in the video information retrieval domain. It is similar to, and adapted, from the text domain where a novel shot is defined as a shot that provides new or previously unseen information on the topic. The benefit of our work is that we have shown, that the detection of novel shots from a list of relevant shots is indeed possible. We considered the various visual and non visual resources and investigated the different resource performances in detecting novel shots.

This research provides a foundation on which additional research can be carried out into the detection of novel shots from a list in the video domain through the

development of an evaluation corpus, including two test collections and ground truth data for the task of novelty detection in the video domain. The fully automatic experiments used to evaluate our novelty detection models allowed us to form conclusions on the effectiveness and benefits of certain feature resources during the detection of novel shots. The rest of this section will briefly review Chapters 1 to 6.

Chapter 1 gave a general overview of information retrieval. We observed that without search engines such as "Google" and "Yahoo!" it is impossible to search through and accurately find all the information available to satisfy our information needs. It discusses the concept of information retrieval, the stages involved, followed by three classical mathematical models to generally describe the information retrieval process. We introduced the need for alternate approaches to the traditional method of information retrieval that currently exist, methods that do not return documents to a users information need, based solely on their degree of relevance alone. We introduced novelty detection as an alternate approach. Novelty detection is defined as the incremental information added to a document based on what the user has already learned from looking at a previous documents in the document list.

Chapter 2 gave a general overview of multimedia information retrieval. We discussed digital video and how it is composed of searchable units. We have seen that the shot is the most common unit of retrieval in video search engines and that the retrieval of video data is much more complex than that of traditional text data. Many challenges exist during the manipulation of video data including the size of the video data itself, the extraction of indexable units and automatic understanding of the semantic meaning from the content displayed in the video sequence. Features detectors have been developed to automatically extract certain features such as the colour and edges contained within the shot.

TREC and TRECVid were also described in Chapter 2. Annually, research groups from all over the world get an opportunity to focus research on specific domains and evaluate the performance of their systems, designed for specific

tasks, using common guidelines and evaluation procedures. TRECVid, a specialised domain of the TREC conferences, was first introduced in 2001 and since then both the tasks and evaluation corpus made available for participants have grown in complexity and size respectively. The tasks for 2003, 2004 and 2005 have focused on T.V news broadcasts. A set of user defined topics were made available each year from which participants could evaluate the performance of their systems.

Chapter 2 outlined the components of a video retrieval system, both the text and image components. We observed that automatic speech transcripts (ASR) are commonly used for video retrieval, however we noted that sometimes ASR transcript words do not accurately define the contents of the particular shots. As a result, it is inadequate to rely on text descriptions alone for the effective retrieval of a relevant shots. Colour, edge and texture can be extracted from a video sequences and these resources can be utilised in a retrieval system in an attempt to aid retrieval performance. High level features and manually annotated semantic features evidences are also used to aid retrieval.

Chapter 3 discussed novelty detection and in particular novelty detection in the text domain. It identified that there are three forms of novelty detection within information retrieval which are closely related however that attempt to accomplish different goals. The first form of novelty detection "event detection" identifies new "event" from across an entire collection of data, where events are defined as "something that happens in some specific time and place" [SC01]. The second kind of novelty detection "Topic tracking" detection focuses on returning new stories on known topics over an entire collection. This thesis focuses on the third kind of novelty detection that of "intra" novelty detection which identifies novel information within a list of shots returned as relevant to a users query, a subset of the collection returned to the users request as opposed to over the entire collection in event detection and topic tracking detection.

Novel data is defined as potentially new data or information not previously seen in any other document so far. While redundant data is defined as data or

information within a document that has been seen by the user already.

Chapter 3 also outlined six assumptions made in order to avoid ambiguity during the identification of novel data. These include

1. Novelty detection is performed on a list of known relevant documents to the user's request.
2. The detection of relevant documents to a user's request is a separate task to the detection of novel documents from a list of relevant documents for a user's request.
3. The novelty of a documents is dependent on the documents that have been previously displayed to the user.
4. Novelty detection is not symmetric.
5. The user is only tolerant of receiving information that he/she may already know due to some background knowledge that he may have on the topic.
6. A user knows nothing about the topic at the time the initial document is displayed and that all knowledge about the topic is gained as a user progresses though a list.

In this Chapter we also described the novelty track which ran as part of the overall TREC conference from 2002 to 2004. The evaluation measures for the novelty track included traditional information retrieval measures namely precision and recall, however in addition the F-measure was used which determines the relative importance of both precision and recall. However it has been noted that the F-measure is not accurate in cross system comparisons as an Fscore can be achieved using a wide variation in both precision and recall values. Also Fscore correlate closely with recall values. This characteristic has been attributed to the consistency of recall values across all topics. It is therefore necessary to also indicate the precision value achieved by the run when referring to its novelty performance.

Finally Chapter 3 introduced and discussed a model "ImportanceValue measure", which we developed for the novelty track in 2004 to accurately detect novel documents from a list of chronologically ordered documents returned from the AQUAINT Collection as relevant to a user's topic. We discussed the performance of the model when compared to other approaches taken in the track using the common set of evaluation measures. The model outperformed all other approaches in the 2004 Novelty track. The model was tested for consistency on the TREC2003 novelty collection and was seen to perform in a similar manner.

Chapter 4 introduced novelty detection in the video domain. It outlined the fact that there is a need for novelty detection models in video collections and in particular within new broadcast collections as overlapping new footage can occur when the collection contains similar stories from more than one broadcaster and also due to the structure of news stories in the form of headlines content body and summary with which, broadcasters present these stories. This can lead to a lot of redundant information occurring in the collection and hence being presented to the searcher of the collection during his/her specific information request.

Video is far more complex to manipulate than text and Chapter 4 outlined the main issues that must be considered during the design of a novelty detection model for the video domain. These include the structure of video, human perception, evolution of news stories and the multiple modalities that can be extracted from a video and used as valuable evidence in video manipulation. The shot is used as the basic unit of manipulation within novelty detection models designed for the video domain. The subjectiveness of what a person perceives as being depicted in an image is an issue within the video domain and in particular for novelty detection within the video domain. It is much more difficult to determines whether a shot is novel when compared to a collection of shots than it is to determine whether a piece of text is novel due to the factual information depicted in text format. As a result novelty models analysing the

visual aspects of the shot, determine a shots novelty on a shot by shot basis.

In theory as news stories evolve, earlier stories on the particular topic will contain a lot more unseen or previously unknown information, while stories occurring later on will not contain as much new information on the topic. As a result shots are ordered chronologically with the oldest shots appearing highest in the list of novel shots.

Chapter 4 also described the various modalities that were utilised for the detection of novel shots. These include text in the form of automatic speech recognition, low level features, including HSV colour, MPEG7 colour structure MPEG7 edge histograms, Canny edge detection evidences and Gabor texture detection evidences; high level features including automatically detected features such as face, anchor person, commercial etc and finally manually annotated concepts presented in MPEG7 descriptions for each shot.

Normalisation is an important part of combining various feature evidences to determine a particular shots novelty value and the normalisation of features for novelty detection was carried out using Histogram normalisation.

The Chapter described the novelty detection models designed for the detection of novel shots from within a results list of shots relevant to a specific user defined topic. The model is broken up into four separate novelty components, namely those utilising text, low level features, automatically detected high level features, and manually annotated concept components. Each component was designed to utilise each of the specific types of feature evidences and determine a shots novelty value based solely on these features. The model was also designed to combine specific features evidences together, to determine a shots novelty value.

To date, in real world applications the most common way to index video for content based retrieval, is by using manually annotated descriptions. These descriptions are provided in the form of a standardised ontology. Chapter 4 described an ontology we have built which is defined specifically for the news

broadcast domain. The ontology was used to annotate the video collection used for our novelty detection experiments.

Chapter 5 discussed the evaluation corpus that was used for the novelty detection experiments. The detection of novel shots from within a chronologically ordered list of known relevant shots for a specific user topic is a new research area within the video domain and it is necessary to create both a test collection that contains a list of relevant shots for each topic, which in this thesis is a subset of the video used in TRECVid 2004 which consisted of news programmes from two different broadcasters, ABC world news tonight and CNN Headline news. It was also necessary to create a corresponding ground truth data collection which contains novel shots, manually assessed for their novelty values by human assessors for each topic in the video test collection.

Two test collections were composed from the results of the best performing search run submission for the manual search task in TRECVid2004. The first collection, Collection_1 consists of shots from the results of the search run that were manually judged as relevant to each specific topic by the NIST assessors. Collection_1 is a subset of Collection_2. Each shots within each topic in Collection_1 was aligned with a story within the original TRECVid2004 collection. Shots within this story judged relevant by NIST assessors for the specific topic were added to Collection_2. The Chapter then discussed the generation of the ground truth data. Four assessors manually judged each shot within each topic to determine whether they were novel or redundant with respect to previously seen shots in the list. An analysis of the ground truth data showed the difference of opinions between assessors on a shot's novelty value, due to the fact that different people perceive what is displayed in an image differently and also due to a persons individual tolerance of redundant data.

Chapter 5 finally discussed the performance evaluation measures which are used to present the results of the novelty detection model experiments, including precision, recall and F-measure.

Chapter 6 presented the results of the automatic experiments carried out to determine the effectiveness of each of the novelty detection models developed for each of the four video resources available including text, low-level feature evidences, high-level feature evidences and manually annotated concept descriptions which were described in Chapter 4. It investigated both the optimal and unbiased F-measure values achieved by each run, over all topics as a whole and on each topic category separately, across Collection_1 and Collection_2 and compared these results to the performance of the baseline run, a system which returned all relevant shots as being novel to each topic.

We conclude that:

- Low level features perform best in our experiments on both collections over all topics and also within each of the individual topic categories. Video content contains various amounts of colour, shapes and motion. As a result it is necessary to use both the colour structure and edge histogram feature evidences available from a videos content, to achieve the best overall performances of novelty detection.
- ASR transcripts do not aid in the detection of novel shots within a list of chronologically ordered shots for specific topic within the video domain.
- Manual concepts show a slight improvement over the baseline novelty performance over some topics, when they are combined with ASR transcripts. We note that the combination of both these features is necessary, due to the fact that many shots are labeled with over used concepts during the manually annotation of video content as a results of the limited number of concepts available in the ontology to describe the contents of the shot. The over usage of particular concepts in the description of shots, leads to many shots being considered redundant, even though they appear visually different. The addition of ASR transcripts reduces this redundancy.
- High level features also performed slightly above the baseline over many topics categories. It was observed however, that high level feature combi-

nations do not perform to the same level of novelty performance achieved by either low level features or the combination of ASR and manual concepts.

7.2 Conclusions

Due to the growth in the television news sector it is becoming necessary to develop "intelligent" methods that determine the novelty value of the information presented.We have seen that typical broadcast TV news programmes contain a wide variety of diverse news topics and these programmes record the evolution of a news story in time containing valuable information for creating documentaries and accessing knowledge on a particular subject. However, we have also seen that collections containing new programmes are rife with repetition as news broadcasters frequently use previously seen video footage on a continuous basis, either in an attempt to remind the viewer of a past story, or as a headline to introduce what is about to be presented within the broadcast, or indeed as a summary of the news programme. Repetition can also occur if a collection contains different news programmes from different broadcasters, as many stories describing the exact same information with perhaps a slight variation of commentary or imagery may be repeated across broadcasters. Traditional video retrieval systems in response to a users query, will return all video sequences which are relevant within a collection, as part of the result list in response to a particular topic, including those that contain exactly the same video footage or graphics already displayed earlier in the results list. This scenario leads to redundant information being displayed to the searcher. As a result, novelty detection in the broadcast news video domain is necessary.

We have seen that novelty detection in the video domain seeks to organise broadcast news search outputs based on the degree of "newness" to the search topic rather than ranking by degree of relevance. Novelty detection techniques have already been applied successfully to the text domain to combat such prob-

lems [Har02, SH03, SH04].

As we have seen, the analysis of video is quite a complex challenge. Video is far more difficult to manipulate than text, mainly due to the fact that, unlike text (where we can attempt to deduce the semantic meaning through words), we have no standard way of extracting the semantic meaning from a video clip. Text spoken during a shot is not a sufficient method of assessing a shot's novelty value as visual content is not aligned with spoken content, this is clearly evident during the commentary of a sports event. It is therefore necessary when manipulating video, to utilise all available resources such as low level feature detection evidences such as colour, edge and texture; high level feature detections evidences such as face, commercial, studio, anchor person and manually annotated concept descriptions. We believe a novelty detection model within the video domain should be broken up into several novelty components capable of incorporating and extracting information from these invaluable resources individually to assess the overall novelty of a shot.

We seen that there is a certain level of subjectiveness inherent in describing what is depicted in an image. This subjectiveness has led to a subjectiveness within the ground truth data, with different assessors having different opinions on a shot's novelty value, based on what they perceive as important in the shot. This was described in more detail in Chapter 5. We observed the method in which an assessor performed the task of novelty detection within the video domain and accurately designed the automatic identification of a shots novelty value, to closely mimic a human being's interaction with the task. We observed that it was necessary, to perform the task on a shot by shot basis rather than against an entire set. It was also necessary to record a decision about a shot's novelty value against a particular shot immediately, before continuing to the next shot in the set. The overall determination of a particular shot's novelty was then based on the accumulation of the shot's novelty value against all shots in the list. If the resulting novelty value was of a sufficient level, then the shot was considered novel, otherwise it was considered redundant. As a result,

novelty detection within the video domain is far more difficult to determine than within the text domain, where novelty detection was carried out against an accumulated list of documents initially.

We seen from the generation of the ground truth data, that the novelty value of a shot is very subjective, as different people have different tolerance levels for the existence of redundant data. Through the analysis of the ground truth data generated by two assessors, we seen that there is a large difference of opinion, between novel and redundant shots. Hence novelty threshold values vary from assessor to assessor. In order to control the amount of novel information that is to exist in a shot, before the shot can be considered a novel, we use threshold values which regulate the level of novel data. The higher the threshold value, θ, the less tolerant the model is to redundant data. This is particularly suited when there is very little tolerance for sifting through shots containing no new information. Decreasing θ decreases the level of novel data which, a shot must contain in order to be considered novel and allows the model to return a greater number of shots as novel. This is more suited to people who don't mind viewing some redundant information in their quest for knowledge on a specific topic.

Due to the fact that novelty detection within the video domain is a new research area it was necessary to develop an evaluation corpus on which to carry out our novelty detection experiments for the determination of the models performance. As a result two new test collections and two corresponding ground truth collections were developed.

We also conclude from the extensive analysis of the novelty models using both ASR as a sole indicator of novelty and also when ASR is combined with other resources, that ASR is not a good feature for detecting novel shots from within a list of shots. We observed that ASR is inconsistent over all topics and in many cases returns all shots as novel or performs worse than the baseline. When we combined ASR with other resources, we observed that in many cases it reduced the performance of the original resource run. It is unclear why ASR does not perform well for novelty detection, however we can guess that it is because the

topics are visual in nature and during the expansion of a shot into its relative news story, the the correlated ASR expansion brings in extra words from the dialogue which are about the news story but not about the visual content. Thus generally they do not bring much value to detecting novel shots. As a result we suggest that ASR should not be solely considered in determining the novelty value of a shot within a topic. We conclude that the detection of novel shots requires the use of other resources available from within the video.

We conclude that low level features are the best resources to use during the detection of novel shots within the a list of relevant shots over each of the topic categories and we note that this is achieved by using the combination of colour structure and edge histograms feature evidences in particular. We believe this is because, colour structure and edge histograms exploit the visual characteristic of the shot which is close in nature to the user's query.

We observed that a number of runs outperform the baseline run, which returns all shots within a list as novel shots. From this we can conclude that the automatic detection of novel shots from within a list of shots within the video domain is indeed possible. We noted not surprisingly, that the manually disambiguated runs outperforms the automatic runs.

We also observed that although our novelty detection models are performing above the baseline over the majority of topics, they are however not achieving the performances of a humans assessors performance of the task. As it is desirable to design a fully automatic novelty detection model which is capable of closely matching the performance of a human interaction with the task, we conclude that there is potential for further research into the area of novelty detection in the video domain.

The work carried out within this research area could have many implications on other research ideas in related and non-related areas including video retrieval search and retrieval and the automatic summarisation of video and multiple videos where the detection of novel or new information is of the up-

most importance for highlighting a movie for example. Novelty detection in general also could be adapted into the research area of information quality.

7.3 Future work

Research into novelty detection models for the detection of novel shots presented in this thesis shows plenty of scope for research to continue into novelty detection models for the video domain and we suggest future work and possible extensions arising out of this work.

Currently the shot is the unit investigation during "intra "novelty detection over broadcast news programmes within the video domain. It would be interesting to investigate intra novelty detection taking stories as the unit of investigation. Stories usually contain a number of shots and as a result will contain a much longer text portions to utilise to determine the novelty of the information being presented. In this case, the interesting thing to observe is that during the determination of a stories novelty using visual features, a sequence of shots keyframes would have to be considered rather than a single specific shot.

Our research into novelty detection has been carried out under the assumption that the user knows nothing about the topic at the time the initial document is displayed and that all knowledge about the topic is gained as a user progresses though a list. We also made the assumption that a user is only tolerant to information that he may already know due to some background knowledge he may have on the topic. It would be interesting to investigate novelty detection taking into account history based profiling for each of the users. This would require the consideration of what the user knows about the topic already. It would require the models to only return novel information on the topic, based what has been previously seen in the topic search results list and also based on what has been seen on this topic previously from other searches, based on the users history profile.

We believe that our text experiments into novelty detection should be performed on other text descriptions available from video such as, closed caption text and optical recognition text, to provide a further understanding of the contributions of text to novelty detection within the video domain. It has been

shown by TREC participants that ASR, closed caption and optical recognition texts combined, perform significantly better than ASR alone.

We believe that novelty detection techniques applied within the video domain should also consider the audio aspect of a video sequence. This would involve creating an evaluation corpus which includes audio. This would however, add significant noise to the data collection and the creation of the ground truth data. However, it would be interesting to observe whether ASR performed well over such a data collection.

Our fusion method for the combination of text and visual component of novelty detection model requires further considerations. In this research fusion is based on a boolean *AND* strategy to determining a shots novelty score, which we believe may have contributed to the poor novelty performance of the combination of text and visual resources in detecting novel shots. A better fusion approach could be identified that may lead to a better novelty performance. We intend to look at different fusion methods including early and late fusion methods, which have been successfully applied within video retrieval approaches.

Another idea would be to investigate a user interactive novelty detection model. It would be interesting to investigate what is the most common level of tolerance for redundant data in a results list, by recoding a users interaction when novelty thresholds can be varied.

Our experiments into the novelty detection, were carried out over a test collection which contained a list of *known* relevant shots to each of the topics. It would be intriguing to investigate the performance of the novelty detection models on an (unjudged or real world) list of results for a specific user defined topic from a traditional retrieval system.

Appendices

Appendix A

TRECVid Topics

Table A.1: TRECVid 2002 search topics.

Topic	Text Description
Topic 75	Find shots with Eddie Rickenbacker in them
Topic 76	Find additional shots with James H. Chandler
Topic 77	Find pictures of George Washington
Topic 78	Find shots with a depiction of Abraham Lincoln
Topic 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
Topic 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.
Topic 81	Find shots of football players
Topic 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.
Topic 83	Find shots of the Golden Gate Bridge
Topic 84	Find shots of Price Tower, designed by Frank Lloyd Wright and built in Bartlesville, Oklahoma,
Topic 85	Find shots containing Washington Square Park's arch in New York City. The entire arch should be visible at some point
Topic 86	Find overhead views of cities - downtown and suburbs. The viewpoint should be higher than the highest building visible
Topic 87	Find shots of oil fields, rigs, derricks, oil drilling/pumping equipment. Shots just of refineries are not desired
Topic 88	Find shots with a map (sketch or graphic) of the continental US.
Topic 89	Find shots of a living butterfly
Topic 90	Find more shots with one or more snow-covered mountain peaks or ridges. Some sky must be visible them behind
Topic 91	Find shots with one or more parrots
Topic 92	Find shots with one or more sailboats, sailing ships, clipper ships, or tall ships - with some sail(s) unfurled
Topic 93	Find shots about live beef or dairy cattle, individual cows or bulls, herds of cattle.
Topic 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).
Topic 95	Find shots of a nuclear explosion with a mushroom cloud
Topic 96	Find additional shots with one or more US flags flapping
Topic 97	Find more shots with microscopic views of living cells
Topic 98	Find shots with a locomotive (and attached railroad cars if any) approaching the viewer
Topic 99	Find shots of a rocket or missile taking off. Simulations are acceptable

Table A.2: TRECVid 2003 search topics.

Topic	Text Description
Topic 100	Find shots with aerial views containing both one or more buildings and one or more roads
Topic 101	Find shots of a basket being made - the basketball passes down through the hoop and net
Topic 102	Find shots from behind the pitcher in a baseball game as he throws a ball that the batter swings at
Topic 103	Find shots of Yasser Arafat
Topic 104	Find shots of an airplane taking off
Topic 105	Find shots of a helicopter in flight or on the ground
Topic 106	Find shots of the Tomb of the Unknown Soldier at Arlington National Cemetery
Topic 107	Find shots of a rocket or missile taking off. Simulations are acceptable
Topic 108	Find shots of the Mercedes logo (star)
Topic 109	Find shots of one or more tanks
Topic 110	Find shots of a person diving into some water
Topic 111	Find shots with a locomotive (and attached railroad cars if any) approaching the viewer
Topic 112	Find shots showing flames
Topic 113	Find more shots with one or more snow-covered mountain peaks or ridges. Some sky must be visible them behind them.
Topic 114	Find shots of Osama Bin Laden
Topic 115	Find shots of one or more roads with lots of vehicles
Topic 116	Find shots of the Sphinx
Topic 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)
Topic 118	Find shots of Congressman Mark Souder
Topic 119	Find shots of Morgan Freeman
Topic 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.
Topic 121	Find shots of a mug or cup of coffee.
Topic 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.
Topic 123	Find shots of Pope John Paul II
Topic 124	Find shots of the front of the White House in the daytime with the fountain running

Table A.3: TRECVid 2004 search topics.

Topic	Text Description
Topic 125	Find shots of a street scene with multiple pedestrians in motion and multiple vehicles in motion somewhere in the shot
Topic 126	Find shots of one or more buildings with flood waters around it/them.
Topic 127	Find shots of one or more people and one or more dogs walking together.
Topic 128	Find shots of U. S. Congressman Henry Hyde's face, whole or part, from any angle.
Topic 129	Find shots zooming in on the U. S. Capitol dome.
Topic 130	Find shots of a hockey rink with at least one of the nets fully visible from some point of view.
Topic 131	Find shots of fingers striking the keys on a keyboard which is at least partially visible.
Topic 132	Find shots of people moving a stretcher.
Topic 133	Find shots of Saddam Hussein.
Topic 134	Find shots of Boris Yeltsin.
Topic 135	Find shots of Sam Donaldson's face - whole or part, from any angle, but including both eyes. No other people visible with him
Topic 136	Find shots of a person hitting a golf ball that then goes into the hole.
Topic 137	Find shots of Benjamin Netanyahu.
Topic 138	Find shots of one or people going up or down some visible steps or stairs.
Topic 139	Find shots of a handheld weapon firing.
Topic 140	Find shots of one or more bicycles rolling along.
Topic 141	Find shots of one or more umbrellas.
Topic 142	Find more shots of a tennis player contacting the ball with his or her tennis racket.
Topic 143	Find shots of one or more wheelchairs. They may be motorized or not.
Topic 144	Find shots of Bill Clinton speaking with at least part of a U. S. flag visible behind him.
Topic 145	Find shots of one or more horses in motion.
Topic 147	Find shots of one or more buildings on fire, with flames and smoke visible.
Topic 148	Find shots of one or more signs or banners carried by people at a march or protest.

Table A.4: TRECVid 2005 search topics.

Topic	Text Description
Topic 149	Find shots of Condoleeza Rice
Topic 150	Find shots of Iyad Allawi, the former prime minister of Iraq
Topic 151	Find shots of Omar Karami, the former prime minister of Lebannon
Topic 152	Find shots of Hu Jintao, president of the People's Republic of China
Topic 153	Find shots of Tony Blair
Topic 154	Find shots of Mahmoud Abbas, also known as Abu Mazen, prime minister of the Palestinian Authority
Topic 155	Find shots of a graphic map of Iraq, location of Bagdhad marked - not a weather map,
Topic 156	Find shots of tennis players on the court - both players visible at same time
Topic 157	Find shots of people shaking hands
Topic 158	Find shots of a helicopter in flight
Topic 159	Find shots of George W. Bush entering or leaving a vehicle (e.g., car, van, airplane, helicopter, etc) (he and vehicle both visible at the same time)
Topic 160	Find shots of something (e.g., vehicle, aircraft, building, etc) on fire with flames and smoke visible
Topic 161	Find shots of people with banners or signs
Topic 162	Find shots of one or more people entering or leaving a building
Topic 163	Find shots of a meeting with a large table and more than two people
Topic 164	Find shots of a ship or boat
Topic 165	Find shots of basketball players on the court
Topic 166	Find shots of one or more palm trees
Topic 167	Find shots of an airplane taking off
Topic 168	Find shots of a road with one or more cars
Topic 169	Find shots of one or more tanks or other military vehicles
Topic 170	Find shots of a tall building (with more than 5 floors above the ground)
Topic 171	Find shots of a goal being made in a soccer match
Topic 172	Find shots of an office setting, i.e., one or more desk tables and one or more computers and one or more people

Appendix B

Ontologies

B.1 206 Ontology

- Program Category
  - Commercial
  - News
  - Entertainment
  - Finance
  - Politics
  - Science/Technology
  - Sports
  - Weather
- Setting/Scene/Site
  - Indoor
    - Studio_Setting
    - Airport_Setting
    - Bank_Setting
    - Church_Setting
    - Court
    - Department_Store_Setting
    - Factory_Setting
    - Hospital_Setting
    - House_Setting
    - Laboratory_Setting
    - MeetingORBoard_Room
    - Night_Club_Setting
    - Office_Setting
    - Press_Conference
    - Restaurant_Setting
    - School_Setting
    - Store_Setting
    - Supermarket_Setting

- Transportation_Setting
- Outdoors
  - Rural_Setting
  - CityScape/Urban_Setting
    - Street_Light
    - City_Street
    - Town_Square
  - Vegetation
    - Flower
    - Tree
    - Forest
    - Greenery
  - Sky
  - Cloud
  - Water_Body
  - Snow
  - Beach
  - Desert
  - Land
  - Mountain
  - Waterfall
  - Bridge
  - Building
    - Building
    - Dome
    - Doorway
    - Ruins
  - Steps_and_Staircases
  - Road
  - Statue_Monumoment
- Outer_Space

People

- Person
  - Adult
    - Female
    - Male
    - Senior_Citizen
  - Juvenile(Child/Teenager)
- Crowd(50+)
- Group(-50)
- Face
- Roles
  - Driver
  - Doctor
  - Nurse
  - Emergency_Services_Personnel
  - Student
  - Teacher
  - Solider
  - Patient
  - Refugee
  - Construction_Worker
  - Pilot
  - Corporate Leader
  - Government Leader/Politican
    - Politican
    - Presidient
    - Prime_Minister
    - Secetary_of_State
  - Military Personnel
  - Police/Private Security
  - Prisoner

- Anchor Person
- Objects
  - Animal
    - Animal
    - Cow
    - Dog
    - Fish
    - Horse
    - Pig
    - Sheep
    - Bird
  - Smoke
  - Barbed_Wire
  - Blackboard
  - Bottle/Drink
  - Camera
  - Candle
  - Chair
  - Clock
  - Crane
  - Drum
  - Flag
    - American_Flag
    - British_Flag
  - Food
  - Handcuffs
  - Keyboard
  - Computer/TV_Screens
  - Microphone
  - Newspaper
  - Parachute

Podium

Sign

Stage

Surfboard

Table

Telephone

Tent

Toy

Tool

Weapon

Gun

Missle

Vehicle

Airplane

Bicycle

Boat/Ship

Bomber_Plane

Bus

Car

Helicopter

Tank

Tractor

Train

Truck

People Activities

Movement

Addressing

Bowing

Carrying

Clapping_Applauding

Cleaning

Crying

Cutting

Cycling

Dancing

Driving

Eating

Embracing

Entering

Fighting

Gesturing

Greeting

Hitting

Interviewing

Kissing

Laughing

Looking_around

Marching

Playing

Posing

Praying

Protesting

Reading

Riding

Shaking_Hands

Shooting

Signing

Singing

Sitting

Skiing

Sleeping

Standing

- Swimming
- Talking/Speaking
- Throwing
- Walking/Running
- Waving

Events

- Explosion/Fire
- Protest
- Natural Disaster
- Sport_Event
  - Sports Event
  - Baseball
  - Basketball
  - Ice_Skating
  - Water
  - Tennis
  - Golf
  - Hockey/Ice-Hockey
  - Snooker
- Transportation_Event
  - Car_Crash
  - Airplane_Takeoff
  - Airplane_Landing
  - Missle_Launch

Graphics

- Charts
- Maps
- Photographs
- Text_Overlay

B.2 Ontology with Descriptions

A. Program Category

1. Commercial: Shots of advertisements, commercials
2. News: Shots depiciting news stories
3. Entertainment: Shots depicting any entertainment segment in action
4. Finance: Shots depicting any finance/business/commerce
5. Politics: Shots depicting any domestic or international politics
6. Science/Technology: Shots depicting any science and technology
7. Sports: Shots depicting any sport in action
8. Weather: Shots depicting any weather related news or bulletin

B. Setting/Scene/Site

1. Indoor: Shots depicting any Indoor Settings
2. Studio_Setting: Shots depicting the interior of a Studio
3. Airport_Setting: Shots depicting any airport
4. Bank_Setting: Shots depicting the interior of a financial bank
5. Church_Setting: Shots depicting the interior of a church
6. Court: Shots depicting the interior of a court
7. Department_Store_Setting: Shots depicting the interior of a department store
8. Factory_Setting: Shots depicting the interior of a factory
9. Hospital_Setting: Shots depicting the interior of a hospital
10. House_Setting: Shots depicting the interior of a home
11. Laboratory_Setting: Shots depicting the interior of a laboratory
12. MeetingORBoard_Room: Shots depicting the interior of a meeting or board room
13. Night_Club_Setting: Shots depicting the interior of a night club
14. Office_Setting: Shots depicting the interior of a office
15. Press_Conference: Shots depicting the interior of a press conference
16. Restaurant_Setting: Shots depicting the interior of a restaurant setting
17. School_Setting: Shots depicting the interior of a school
18. Store_Setting: Shots depicting the interior of a store

19. Supermarket_Setting: Shots depicting the interior of a supermarket

20. Transportation_Setting: Shots depicting the interior of a bus station/train station transportation setting

21. Outdoors: Shots depicting Outdoor settings

22. Rural_Setting: Shots depicting any rural setting

23. CityScape/Urban_Setting: Shots depicting any CityScape or Urban setting

24. Street_Light: Shots depicting any a street light

25. City_Street: Shots depicting any city street

26. Town_Square: Shots depicting any town square

27. Vegetation: Shots depicting any natural vegetation either in foreground or backgound

28. Flower: Shots depicting any flower either in foreground or backgound

29. Tree: Shots depicting any tree either in foreground or backgound

30. Forest: Shots depicting any forest either in foreground or backgound

31. Greenery: Shots depicting any greenary such as grass or hedges either in foreground or backgound

32. Sky: Shots depicting the sky either in foreground or backgound

33. Cloud: Shots depicting a cloud either in foreground or backgound

34. Water_Body: Shots depicting any lake, river , sea either in foreground or backgound

35. Snow: Shots depicting any snow either in foreground or backgound

36. Beach: Shots depicting any beach either in foreground or backgound

37. Desert: Shots depicting any desert either in foreground or backgound

38. Land: Shots depicting any land mass

39. Mountain: Shots depicting any mountain or mountain range with the slopes visible

40. Waterfall: Shots depicting a waterfall

41. Bridge: Shots depicting any bridge

42. Building: Shots depicting the exterior of any building

43. Dome: Shots depicting the exterior of a dome

44. Doorway: Shots depicting the exterior of a doorway

45. Ruins: Shots depicting the exterior of a ruin

46. Steps_and_Staircases: Shots depicting any steps or stairways

47. Road: Shots depicting a road

48. Statue_Monumoment: Shots depicting a statue or monument

49. Outer_Space: Shots depicting outerspace

C. People

1. Person: Shots depicting any person

2. Adult: Shots depicting an adult

3. Female: Shots depicting a female

4. Male: Shots depicting a male

5. Senior_Citizen: Shots depicting a senior citizen

6. Juvenile(Child/Teenager): Shots depicting a child or teenager

7. Crowd(50+): Shots depicting a crowd of fifty or more people

8. Group(−50): Shots depicting a group of up to fifty people

9. Face: Shots depicting a face

10. Roles

11. Driver: Shots depicting a driver

12. Doctor: Shots depicting a doctor in medical profession

13. Nurse: Shots depicting a nurse

14. Emergency_Services_Personnel: Shots depicting any personal in the emergency service occupation

15. Student: Shots depicting any students

16. Teacher: Shots depicting teachers

17. Solider: Shots depicting a solider

18. Patient: Shots depicting a patient

19. Refugee: Shots depicting a refugee

20. Construction_Worker: Shots depicting construction workers

21. Pilot: Shots depicting a pilot

22. Corporate Leader: Shots depicting any person who is a corporate leader e.g. CEO, CFO, Managing Director, Media Manager etc.

23. Government Leader/Politican: Shots depicting any person who is a gov-

erning leader

24. Politican: Shots depicting any politican

25. Presidient: Shots depicting any president of a country

26. Prime_Minister: Shots depicting any country

27. Secetary_of_State: Shots depicting the secetary of state

28. Military Personnel: Shots depicting any military personnel

29. Police/Private Security: Shots depicting any law enforcement or private security agency personnel

30. Prisoner: Shots depicting any person imprisoned, behind bars, in jail or in handcuffs

31. Anchor Person: Shots depicting an anchor person in broadcast news

D. Objects

1.Animal: Shots depicting any animal

2. Cow: Shots depicting any cow

3. Dog: Shots depicting any dog

4. Fish: Shots depicting fish

5. Horse: Shots depicting a horse

6. Pig: Shots depicting a pig

7. Sheep: Shots depicting a sheep

8. Bird: Shots depicting a bird

9. Smoke: Shots depicting any smoke

10. Barbed_Wire: Shots depicting any basbed wire

11. Blackboard: Shots depicting any blackboard

12. Bottle/Drink: Shots depicting any bottle/drink

13. Camera: Shots depicting any camera

14. Candle: Shots depicting any candle

15. Chair: Shots depicting any chair

16. Clock: Shots depicting any clock

17. Crane: Shots depicting any building crane on a building site

18. Drum: Shots depicting any drum

19. Flag: Shots depicting any flag

20. American_Flag: Shots depicting the US flag
21. British_Flag: Shots depicting the British flag
22. Food: Shots depicting any food
23. Handcuffs: Shots depicting handcuffs
24. Keyboard: Shots depicting a keyboard
25. Computer/TV_Screens: Shots depicting a TV or computer screen
26. Microphone: Shots depicting a microphone
27. Newspaper: Shots depicting a newspaper
28. Parachute: Shots depicting a parachute
29. Podium: Shots depicting a podium
30. Sign: Shots depicting a sign
31. Stage: Shots depicting a stage
32. Surfboard: Shots depicting a surfboard
33. Table: Shots depicting a table
34. Telephone: Shots depicting a telephone
35. Tent: Shots depicting a tent
36. Toy: Shots depicting a toy
37. Tool: Shots depicting a piece of equipment or tool
38. Weapon: Shots depicting any weapon
39. Gun: Shots depicting a gun
40. Missle: Shots depicting a missle
41. Vehicle: Shots depicting any vehicle
42. Airplane: Shots depicting any airplane
43. Bicycle: Shots depicting any bicycle
44. Boat/Ship: Shots depicting any boat/ship
45. Bomber_Plane: Shots depicting any war plane or bomber plane
46. Bus: Shots depicting any bus
47. Car: Shots depicting any car
48. Helicopter: Shots depicting any helicopter
49. Tank: Shots depicting any tank
50. Tractor: Shots depicting any tractor

51. Train: Shots depicting any train
52. Truck: Shots depicting any truck

E.People Activities Movement

1. Addressing: Shots depicting a person addressing a person or group of people
2. Bowing: Shots depicting a person bowing
3. Carrying: Shots depicting a person carrying something
4. Clapping_Applauding: Shots depicting a person clapping or applauding
5. Cleaning: Shots depicting a person cleaning something
6. Crying: Shots depicting a person crying
7. Cutting: Shots depicting a person cutting something
8. Cycling: Shots depicting a person cycling
9. Dancing: Shots depicting a person dancing
10. Driving: Shots depicting a person driving
11. Eating: Shots depicting a person eating
12. Embracing: Shots depicting a person embracing something or someone
13. Entering: Shots depicting a person entering a room
14. Fighting: Shots depicting a person fighting with someone or group of people
15. Gesturing: Shots depicting a person gesturing
16. Greeting: Shots depicting a person greeting someone or group of people
17. Hitting: Shots depicting a person hitting something or someone
18. Interviewing: Shots depicting a person interviewing someone
19. Kissing: Shots depicting a person kissing someone or something
20. Laughing: Shots depicting a person laughing
21. Looking_around: Shots depicting a person looking around them
22. Marching: Shots depicting a person marching in a parade or protest
23. Playing: Shots depicting a person playing
24. Posing: Shots depicting a person posing
25. Praying: Shots depicting a person praying
26. Protesting: Shots depicting a person protesting
27. Reading: Shots depicting a person reading
28. Riding: Shots depicting a person riding a horse or bicycle

29. Shaking_Hands: Shots depicting a person shaking hands with someone

30. Shooting: Shots depicting a person shooting

31. Signing: Shots depicting a person signing

32. Singing: Shots depicting a person singing

33. Sitting: Shots depicting a person sitiing

34. Skiing: Shots depicting a person skiing

35. Sleeping: Shots depicting a person sleeping

36. Standing: Shots depicting a person standing

37. Swimming: Shots depicting a person swimming

38. Talking/Speaking: Shots depicting a person talking or speaking

39. Throwing: Shots depicting a person throwing something

40. Walking/Running: Shots depicting a person walking or running

41. Waving: Shots depicting a person waving

F. Events

1. Explosion/Fire: Shots depicting any explosion or fire

2. Protest: Shots depicting a protest

3. Natural Disaster: Shots depicting any the aftermaths of a natural disaster such as a flood, hurricane, earthquake

4. Sport_Event: Shots depicting any sports event

5. Baseball: Shots depicting any baseball game

6. Basketball: Shots depicting any basketball match

7. Ice_Skating: Shots depicting ice skating

8. Water: Shots depicting any water sports such as water skiing

9. Tennis: Shots depicting any tennis match

10. Golf: Shots depicting any game of golf

11. Hockey/Ice-Hockey: Shots depicting any hockey match

12. Snooker: Shots depicting any snooker match

13. Transportation_Event: Shots depicting any transportation event

14. Car_Crash: Shots depicting any car crash

15 Airplane_Takeoff: Shots depicting any airplane taking off

16. Airplane_Landing: Shots depicting any airplane landing

17. Missle_Launch: Shots depicting any missle launching

G. Graphics

1. Charts: Shots depicting any charts
2. Maps: Shots depicting any maps
3. Phtotographs: Shots depicting any photgraphs
4. Text_Overlay: Shots depicting any text overlay

B.3 LSCOM-Lite Ontology with Descriptions

The following is a list of the LSCOM-lite ontology including a description for each concept which were provided by LSCOM [gui].

A. Program Category

1. Politics: News items about domestic or international politics
2. Finance/Business: News items about finance/business/commerce
3. Science/Technology: News items about science and technology
4. Sports: Shots depicting any sport in action
5. Entertainment: Shots depicting any entertainment segment in action
6. Weather: Shots depicting any weather related news or bulletin
7. Commercial/Advertisement: Shots of advertisements, commercials

B. Setting/Scene/Site

1. Indoor: Shots of Indoor locations
2. Court: Shots of the interior of a court-room location
3. Office: Shots of the interior of an Office Setting
4. Meeting: Shots of a Meeting taking place indoors
5. Studio Setting: Shots of the studio setting including anchors, interviews and all events that happen in a news room
6. Outdoor: Shots of Outdoor locations
7. Building: Shots of an exterior of a building
8. Desert: Shots with the desert in the background
9. Vegetation: Shots depicting natural or artificial greenery, vegetation woods, etc.
10. Mountain: Shots depicting a mountain or mountain range with the slopes visible
11. Road: Shots depicting a road
12. Sky: Shots depicting sky
13. Snow: Shots depicting snow

14. Urban-Setting: Shots depicting an urban or suburban setting

15. Waterscape/Waterfront: Shots depicting a waterscape or waterfront

C. People

1. Crowd: Shots depicting a crowd

2. Face: Shots depicting a face

3. Person: Shots depicting a person. The face may be partially visible

Roles

4. Government Leader: Shots of a person who is a governing leader e.g. president, prime-minister, chancellor of the exchequer, etc.

5. Corporate Leader: Shots of a person who is a corporate leader e.g. CEO, CFO, Managing Director, Media Manager etc.

6. Police/Private Security Personnel: Shots depicting law enforcement or private security agency personnel

7. Military: Shots depicting the military personnel

8. Prisoner: Shots depicting a person imprisoned, behind bars, in jail or in handcuffs

D. Objects

1. Animal (No humans): Shots depicting an animal.

2. Computer or Television Screens: Shots depicting television or computer screens

3. Flag-US: Shots depicting a US flag Vehicle

4. Airplane: Shots of an airplane

5. Car: Shots of a car

6. Bus: Shots of a bus

7. Truck: Shots of a truck

8. Boat/Ship: Shots of a boat or ship

E. People Activities Movements

1. Walking/Running: Shots depicting a person walking or running

2. Parade: Shots depicting a parade with people marching

F. Events

1. Explosion/Fire: Shots of an explosion or a fire

2. Protest: People marching with banners, flags, posters

3. Natural Disaster: Shots depicting the happening or aftermath of a natural disaster such as earthquake, flood, hurricane, tornado, tsunami

G. Graphics

1. Maps: Shots depicting regional territory graphically as a geographical or political map

2. Charts: Shots depicting any graphics that is artificially generated such as bar graphs, line charts etc. Maps should not be included

Appendix C

Assessor Guidelines

Novelty Experiments Assessor's Guidelines

Given a chronologically ordered list of known relevant shots to a particular topic, reduce this list to contain only shots that provide novel information on the topic while at the same time maintaining the original list ordering.

The order of the shots is important in this experiment.

Note:

1. The first shot in the set is ALWAYS novel, as it is assumed that you know nothing about the topic at the time the initial document is displayed and that all knowledge about the topic is gained as you progress through a list.
2. The topic is very important in this analysis. The assessor is asked to refer back to the topic during the assessment of a shot's novelty value for each topic.

Instruction to Assessors:

1. Read the topic.
2. Place the first shot into the novel set.
3. Go through the list and compare each new shot with the shots already present in the novel set. Continuously refer to the topic.
4. If, in your opinion, the current shot contains absolutely no new information compared to shots you have previously seen then this shot should be placed in the redundant set.
5. If, in your opinion, the current shot contains insignificant amount of new information and for the most part contains a high level of redundant information (adding nothing new to the knowledge you have already gained) when compared to shots already present in the novel set then place this shot in the redundant set otherwise place it in the novel set.
6. If the current shot contains new information compared to shots already present in the novel set then place this shot in the novel set.
7. Continue the process for each subsequent shot in the chronologically ordered list for all topics in the video collection.

Appendix D

Experimental Run Threshold Values

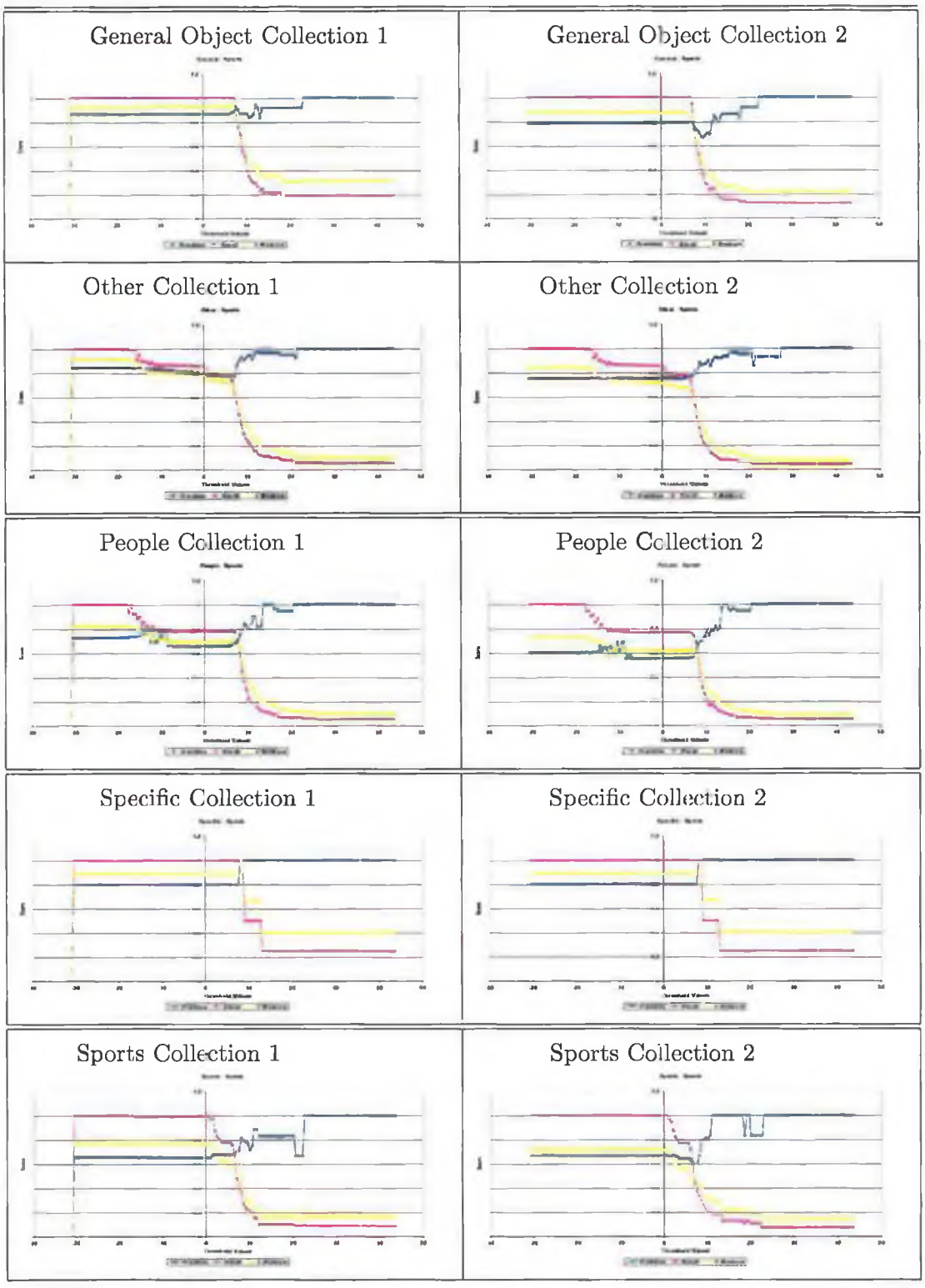

Figure D.1: Threshold variation graphs over the both Collection_1 and Collection_2 for the high level feature "Sports" run

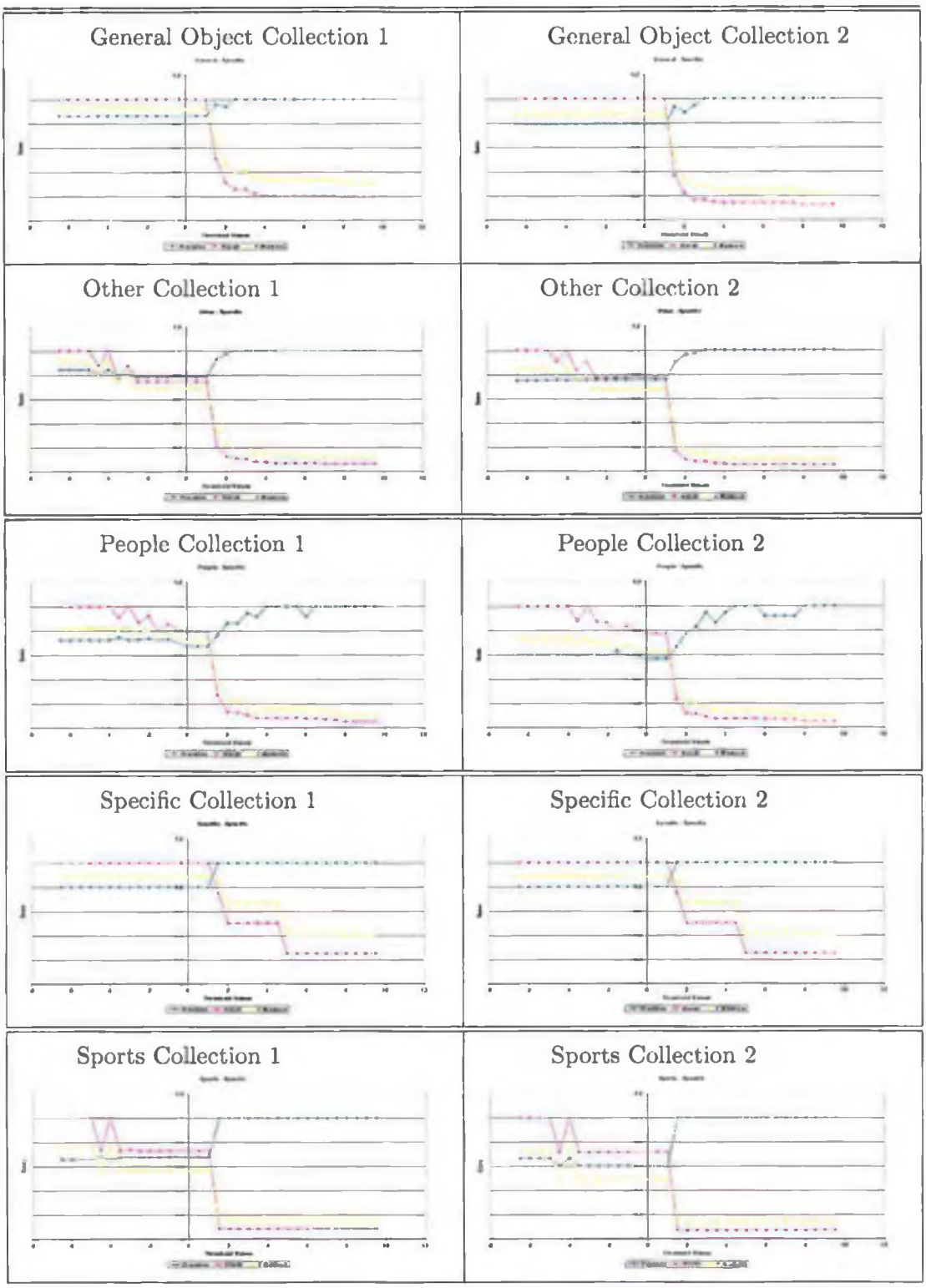

Figure D.2: Threshold variation graphs over the both Collection_1 and Collection_2 for the high level feature "Specific" run

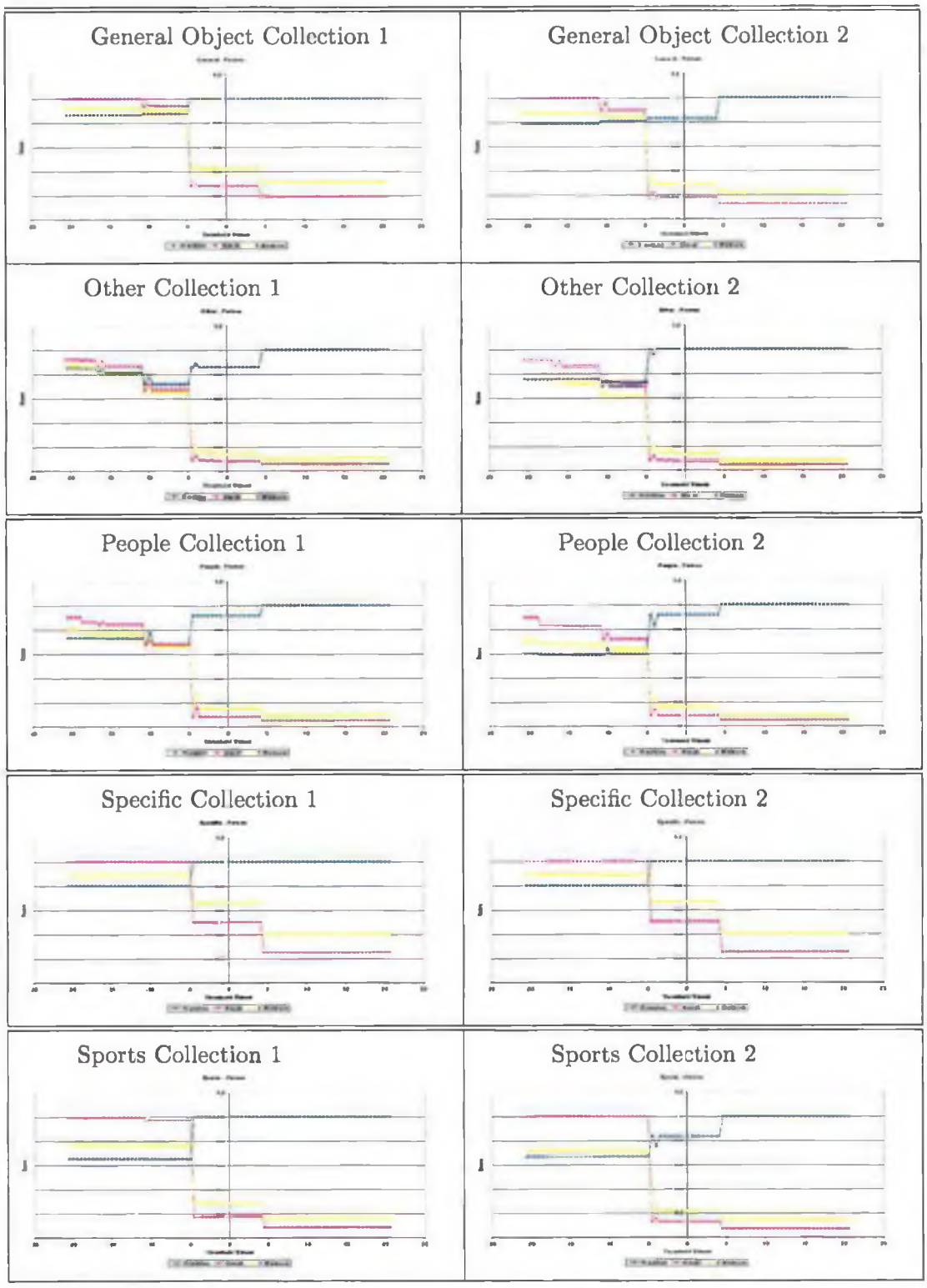

Figure D.3: Threshold variation graphs over the both Collection_1 and Collection_2 for high level features "People" run

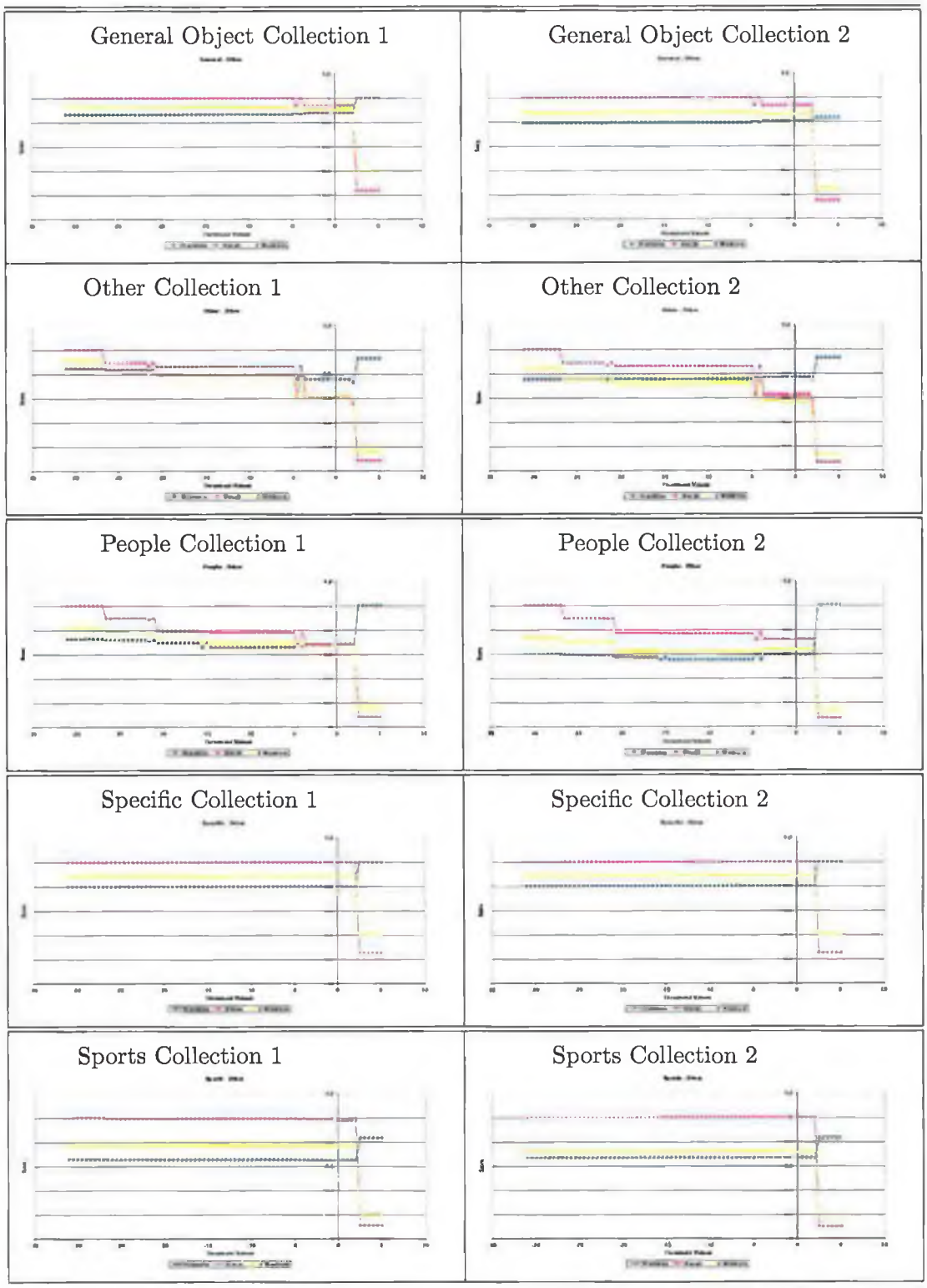

Figure D.4: Threshold variation graphs over the both Collection_1 and Collection_2 for high level features "Other" run

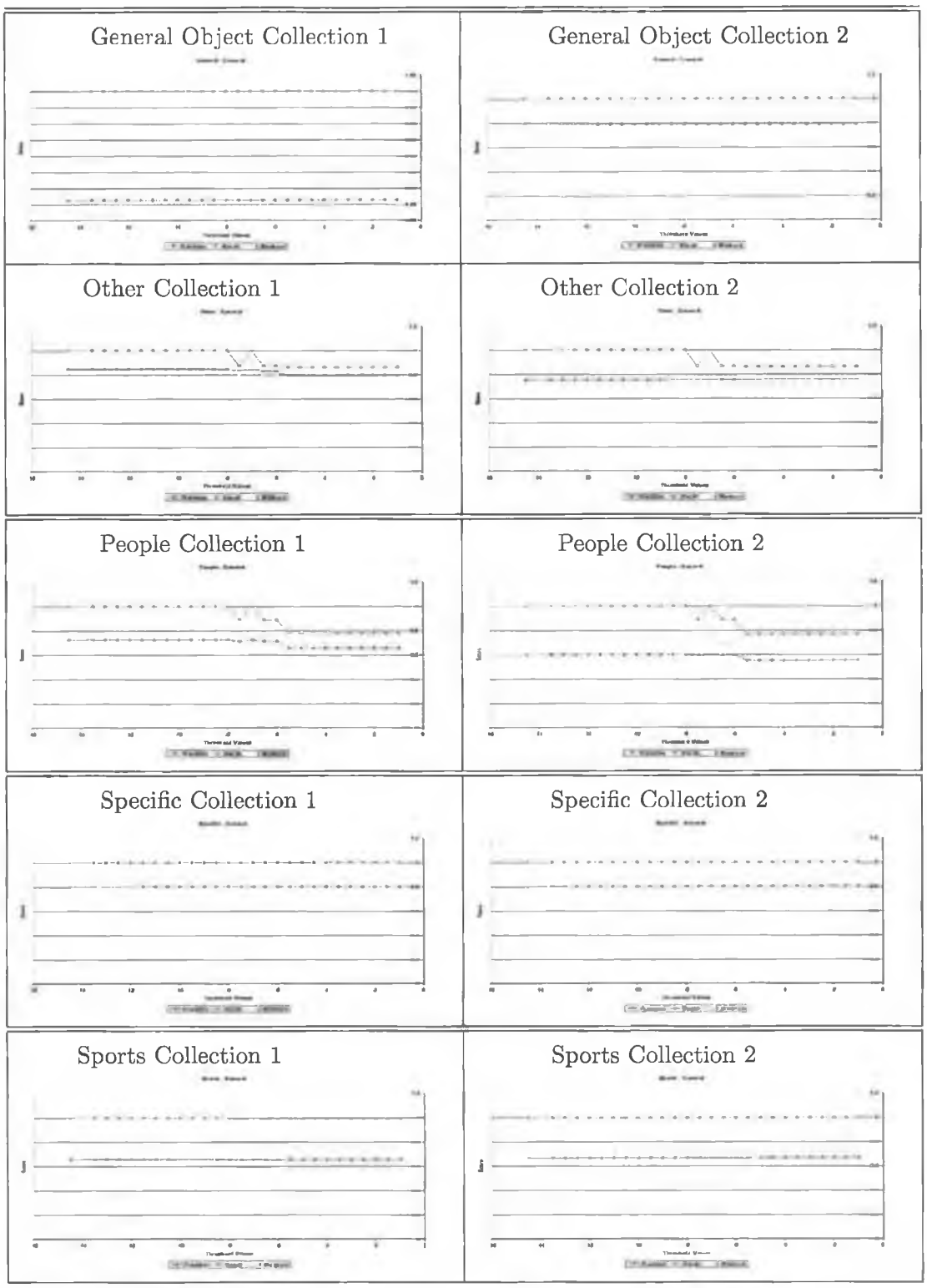

Figure D.5: Threshold variation graphs over the both Collection_1 and Collection_2 for high level features "General" run

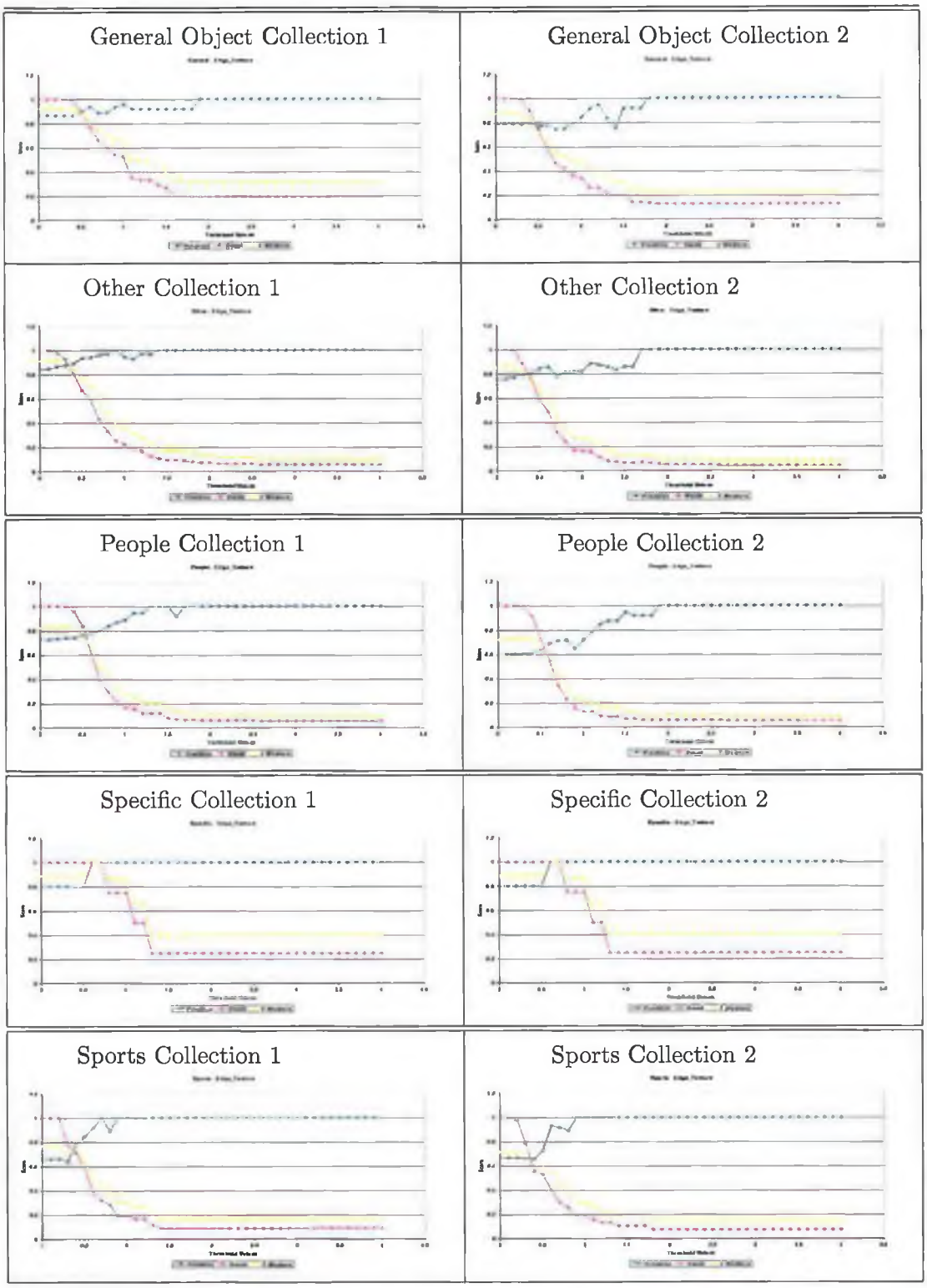

Figure D.6: Threshold variation graphs over the both Collection_1 and Collection_2 for low level features "Edge_Texture" run

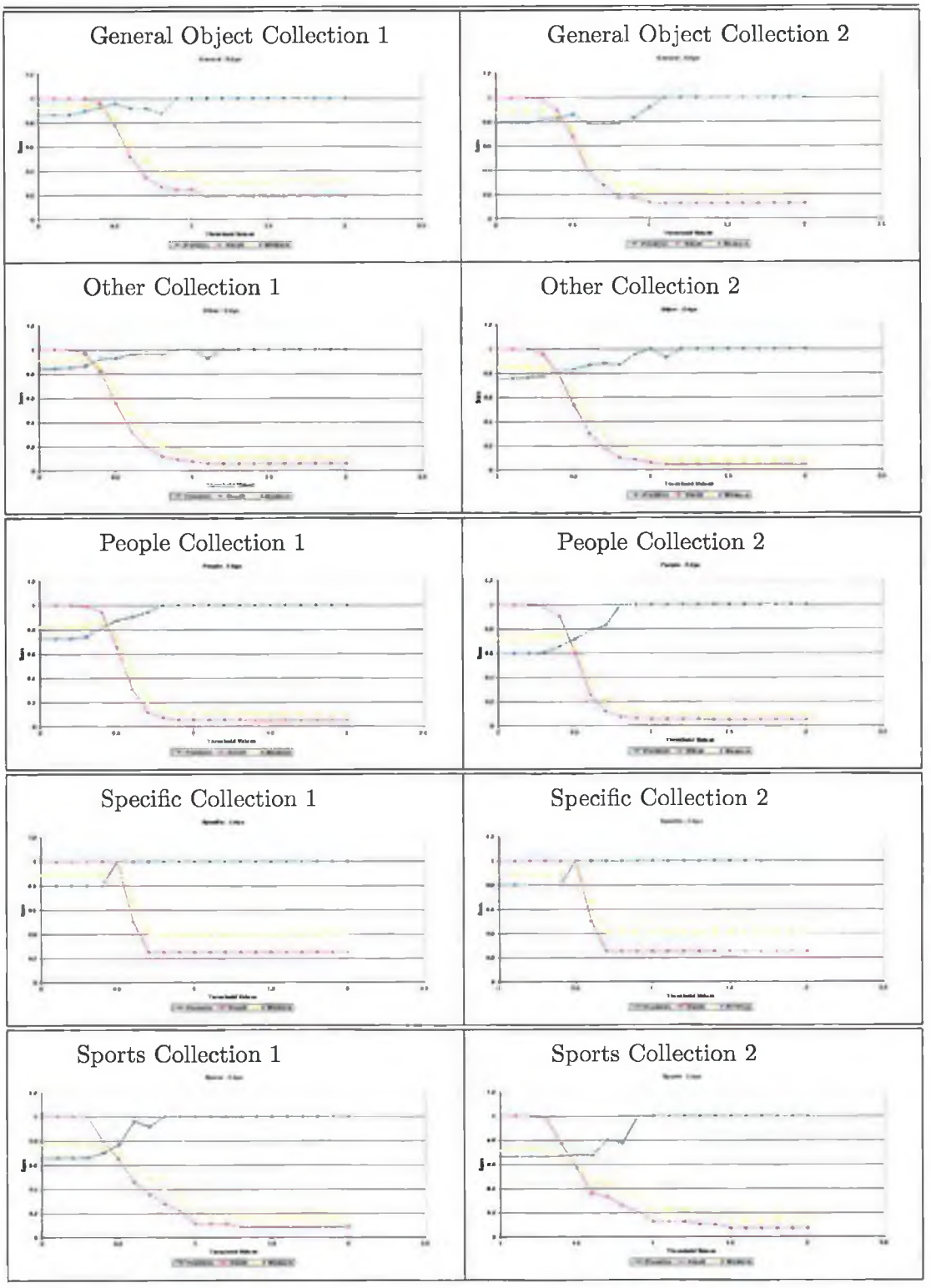

Figure D.7: Threshold variation graphs over the both Collection_1 and Collection_2 for low level features "EdgeHist" run

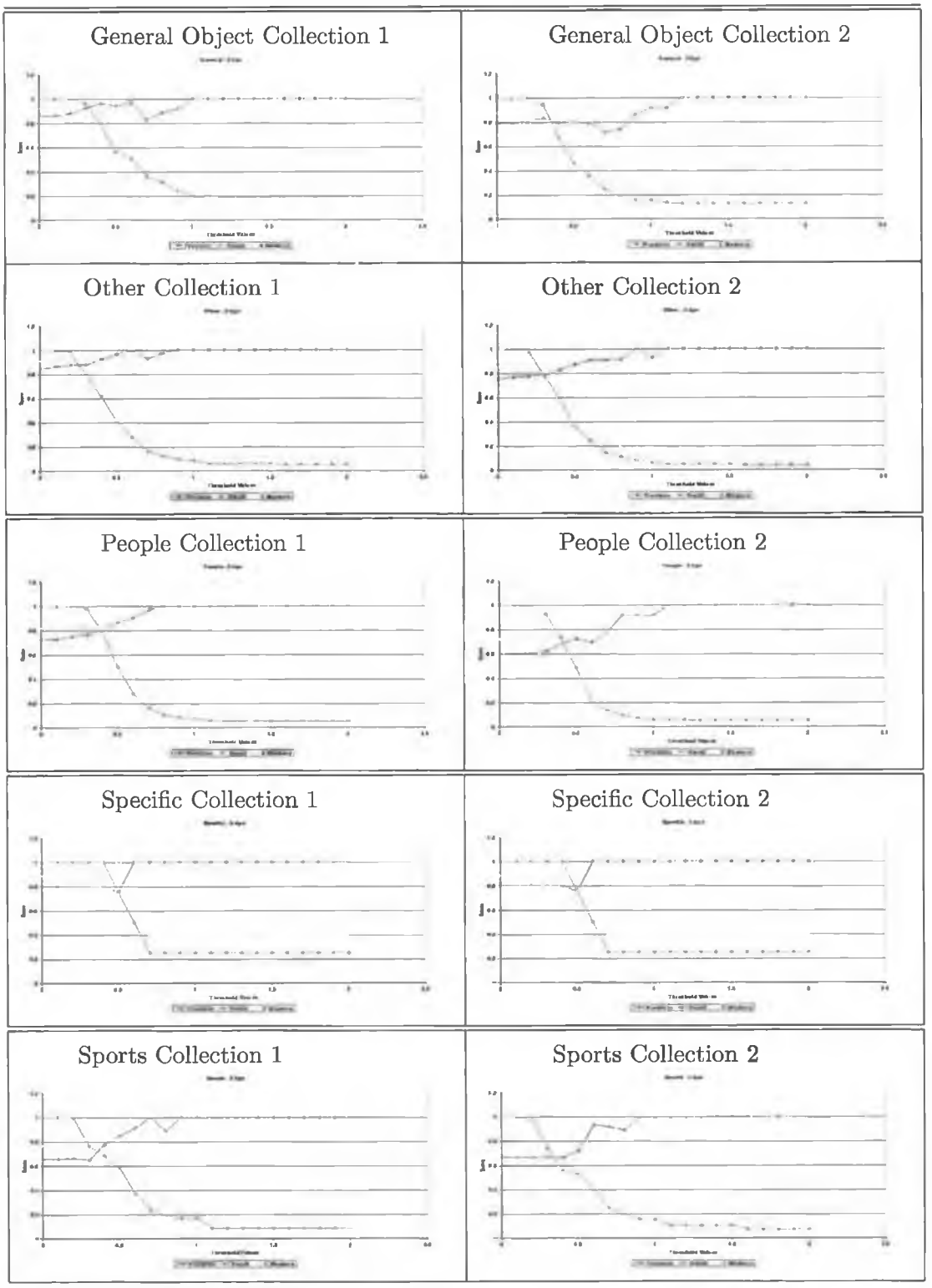

Figure D.8: Threshold variation graphs over the both Collection_1 and Collection_2 for low level features "Canny edge" run

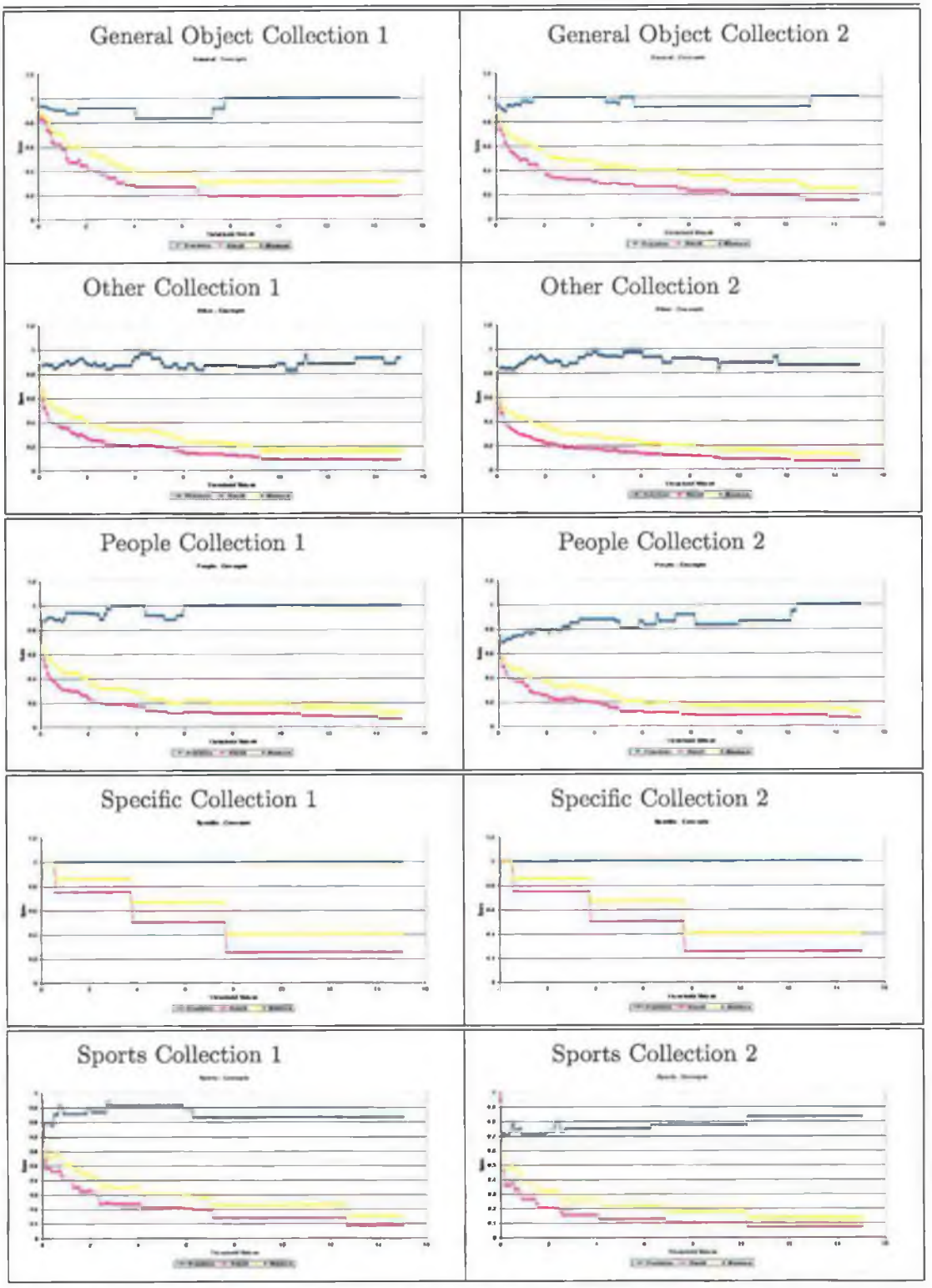

Figure D.9: Threshold variation graphs over the both Collection_1 and Collection_2 for manually annotated concepts "Concepts" run

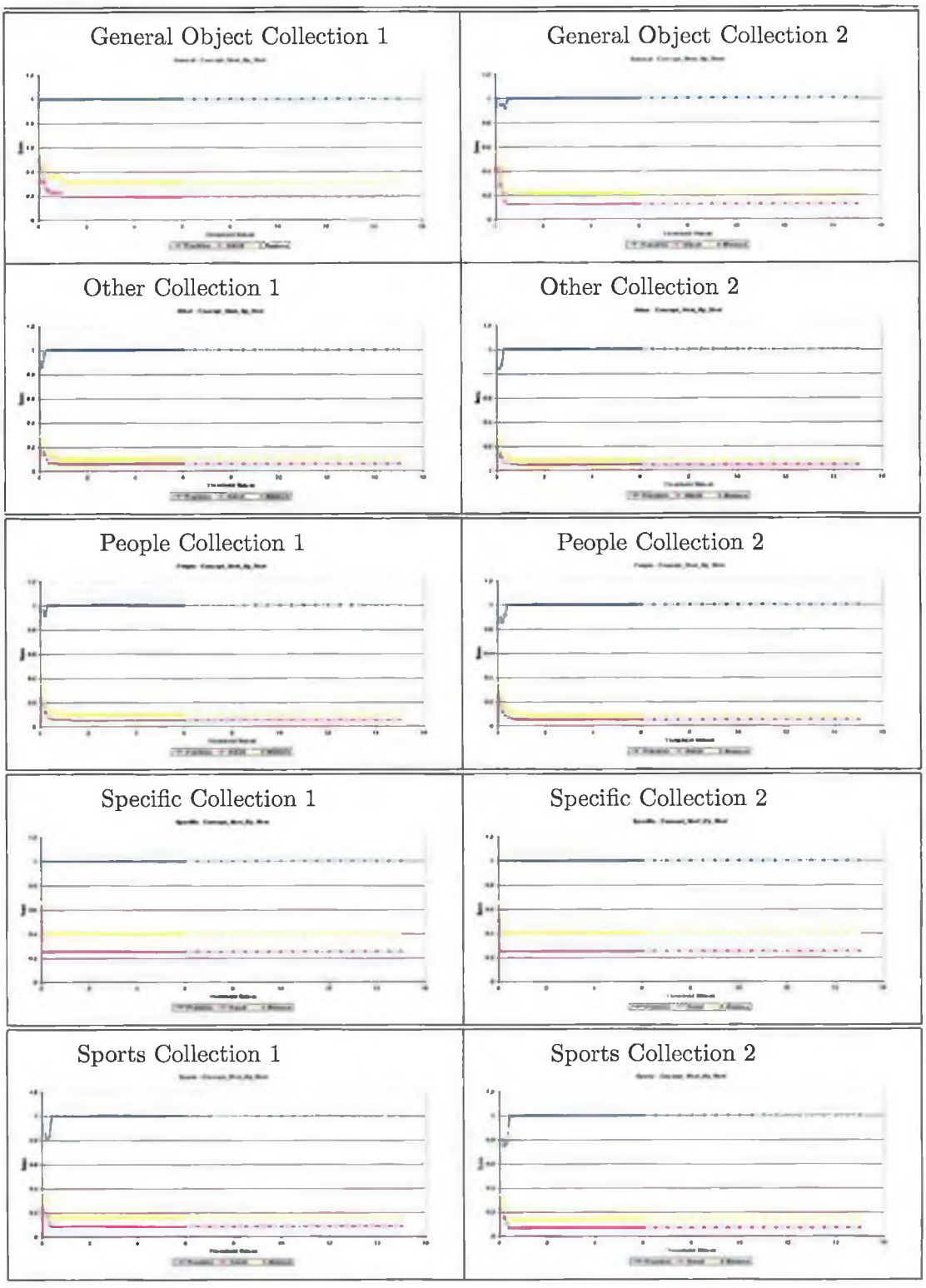

Figure D.10: Threshold variation graphs over the both Collection_1 and Collection_2 for manually annotated concepts "ASR_concept" run

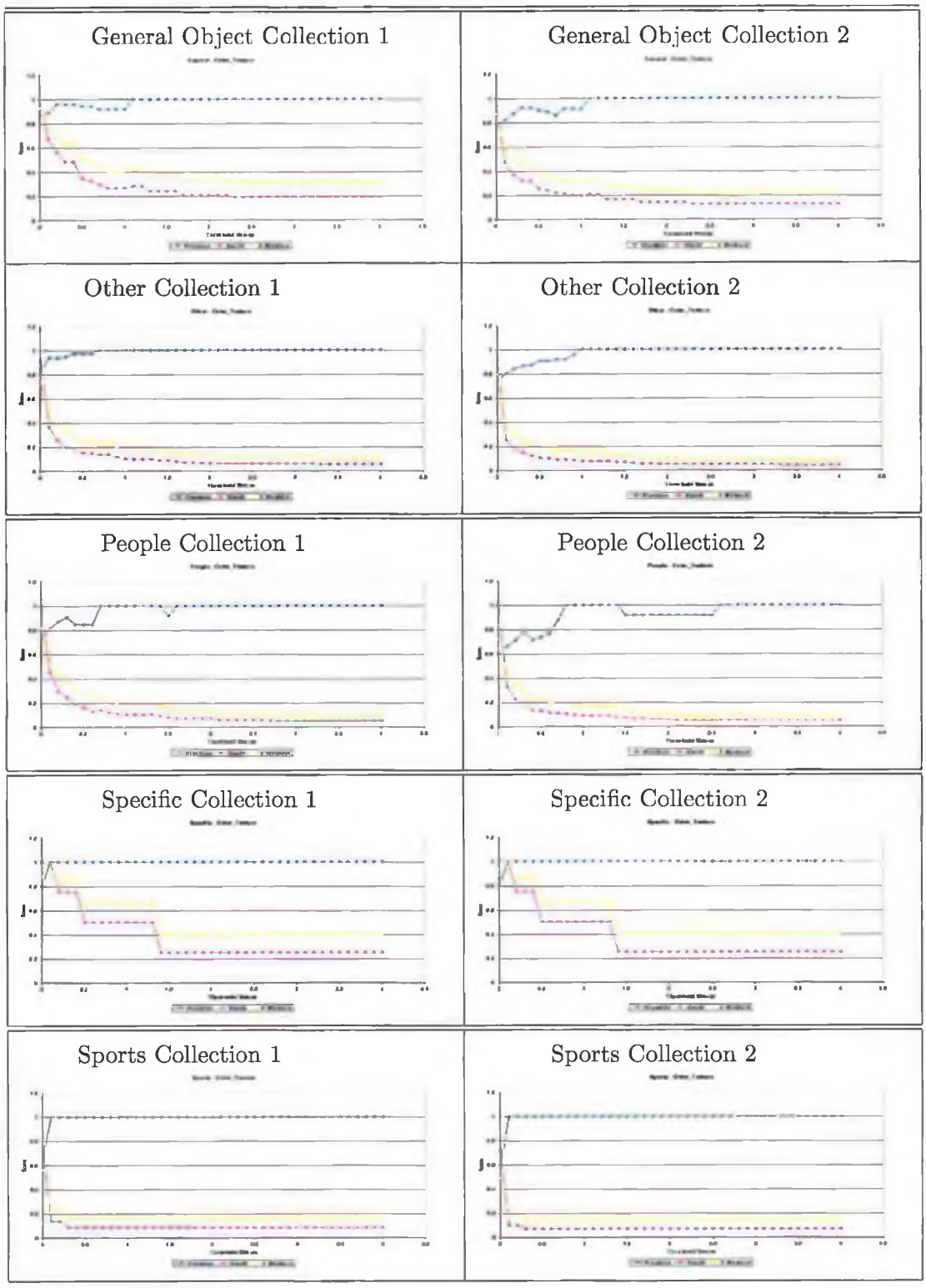

Figure D.11: Threshold variation graphs over the both Collection_1 and Collection_2 for low level features "HSVColor_Texture" run

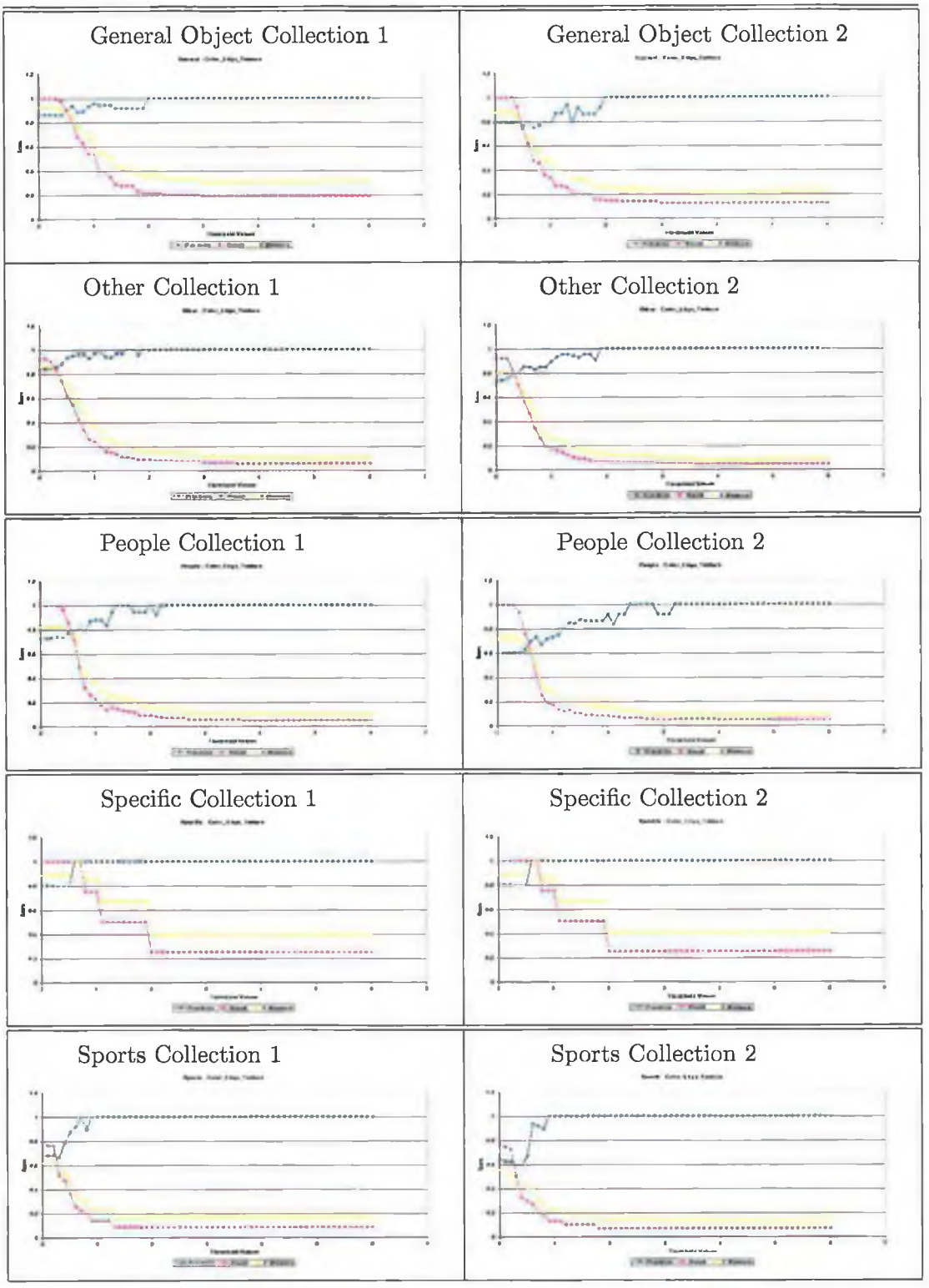

Figure D.12: Threshold variation graphs over the both Collection_1 and Collection_2 for low level features "HSVColor_CannyEd_Texture" run

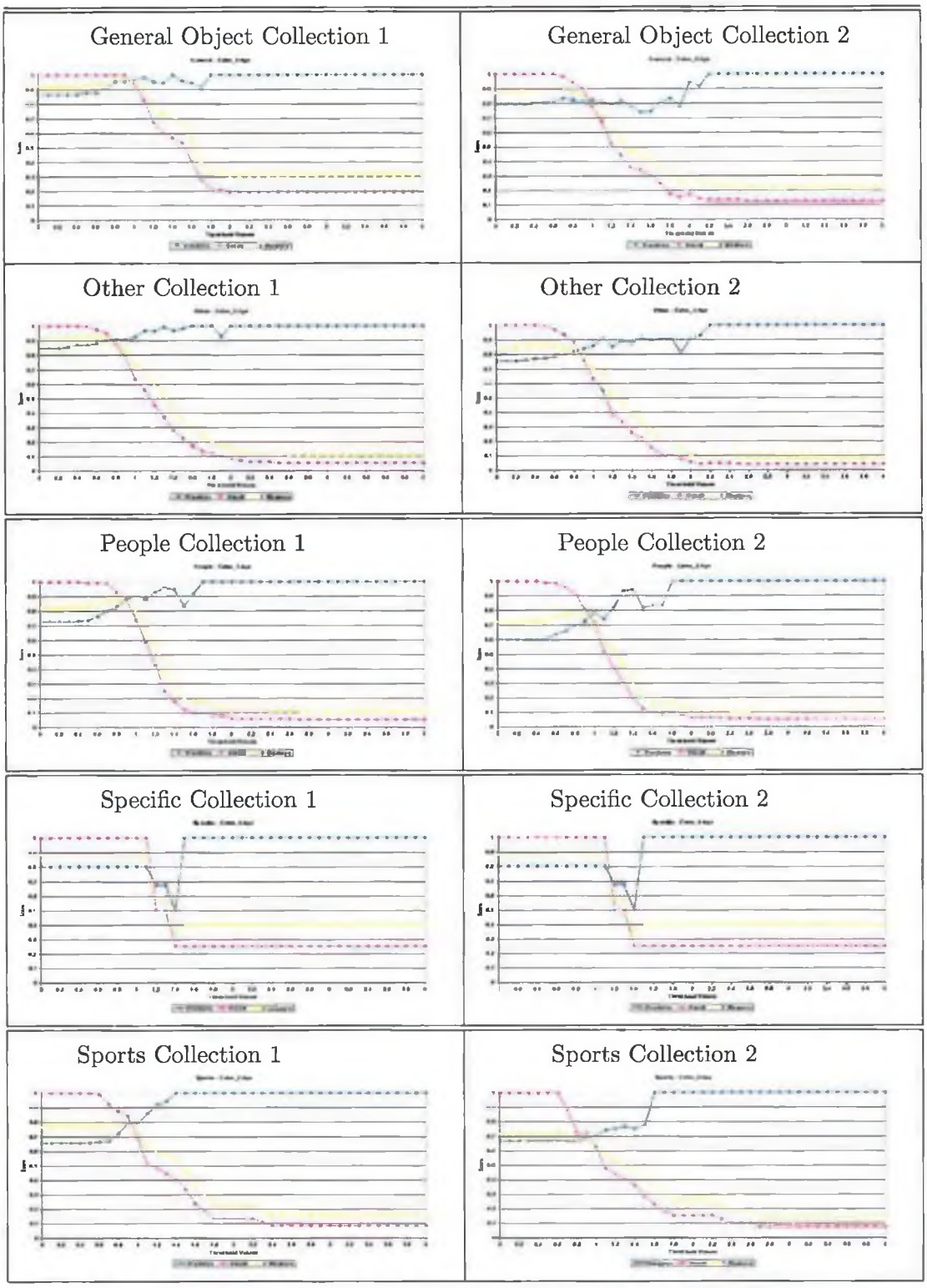

Figure D.13: Threshold variation graphs over the both Collection_1 and Collection_2 for low level features "ColorStruc_EdgeHist" run

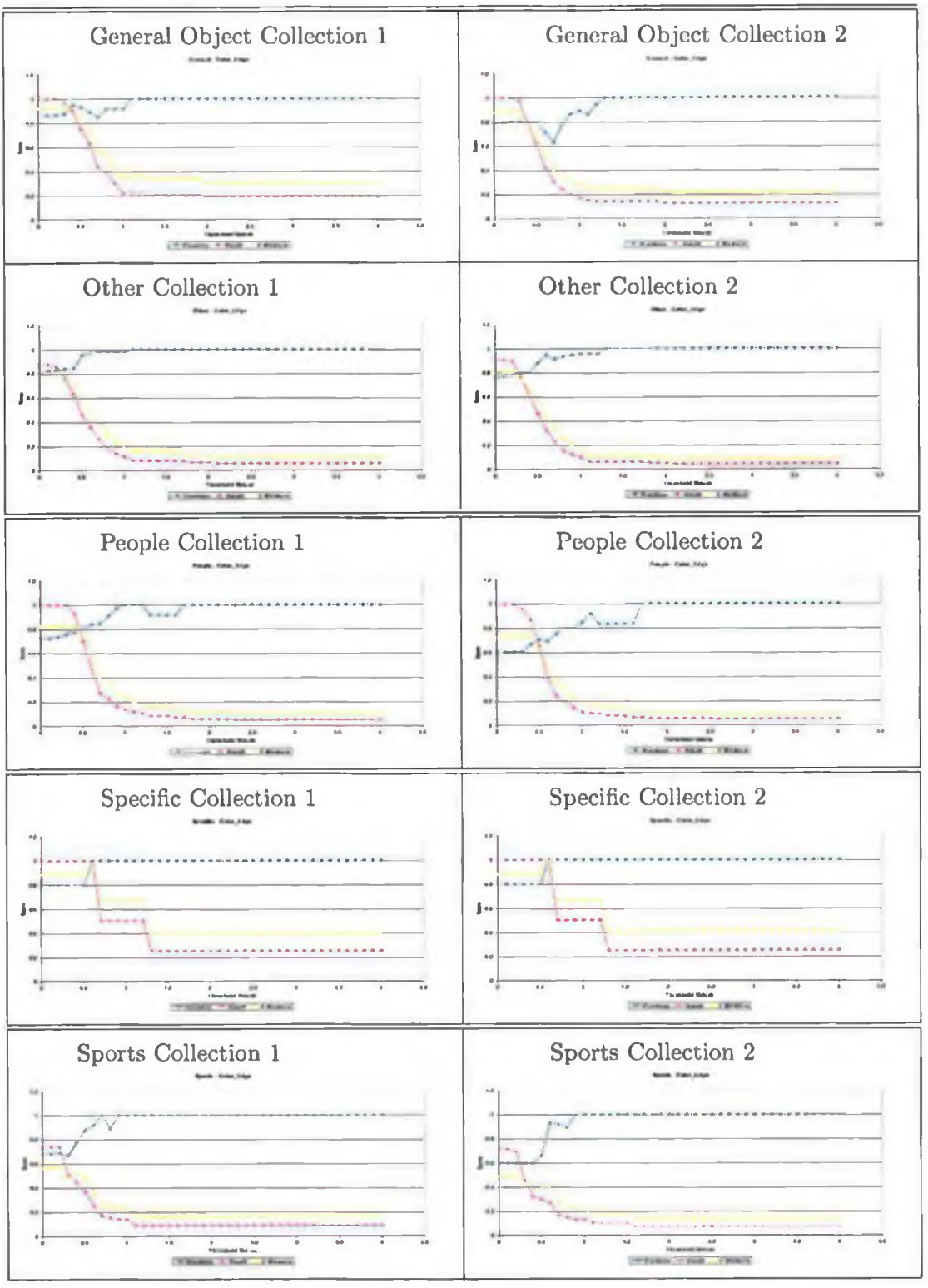

Figure D.14: Threshold variation graphs over the both Collection_1 and Collection_2 for low level features "HSVColor_CannyEd" run

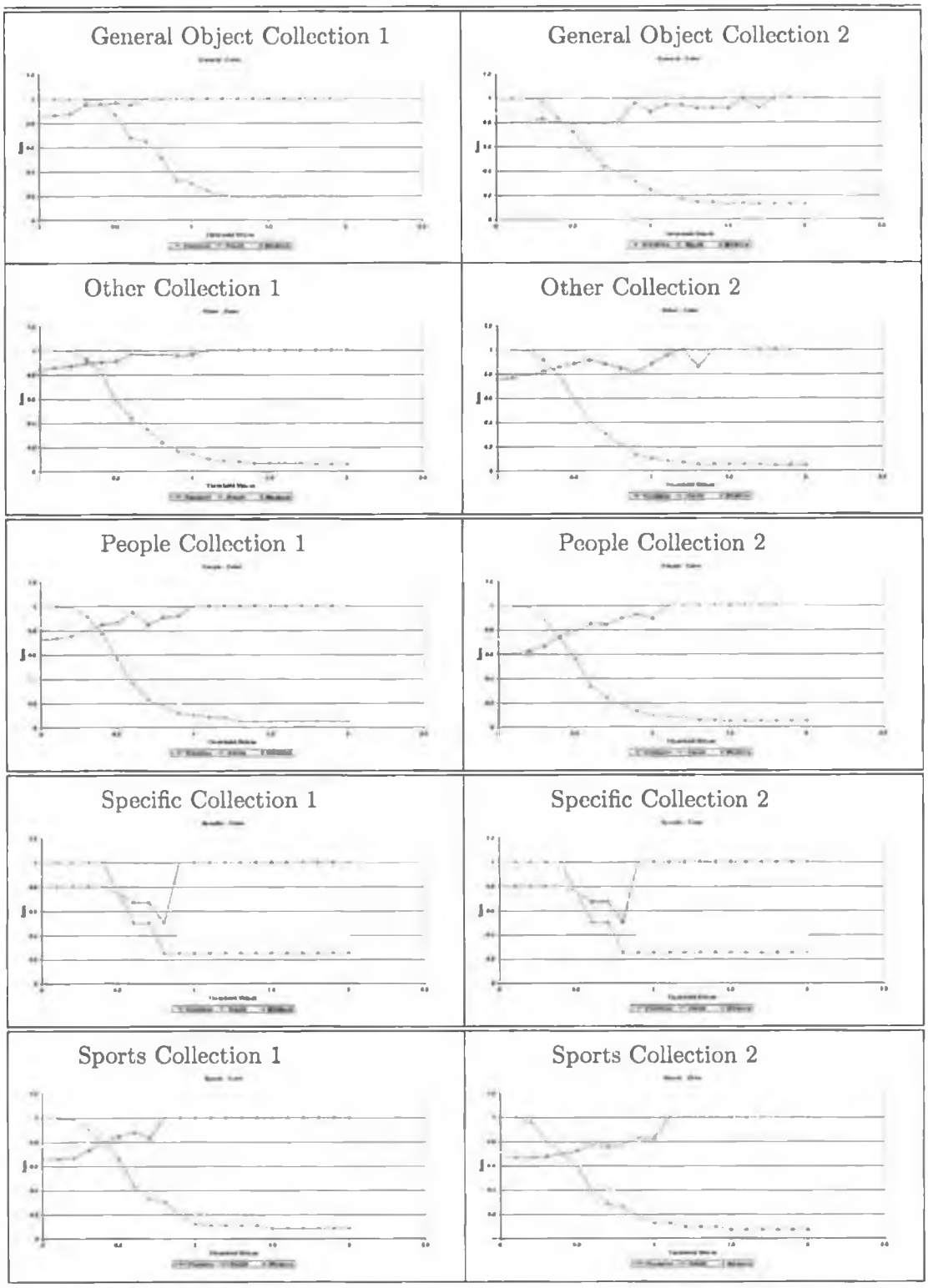

Figure D.15: Threshold variation graphs over the both Collection_1 and Collection_2 for low level features "ColorStruc" run

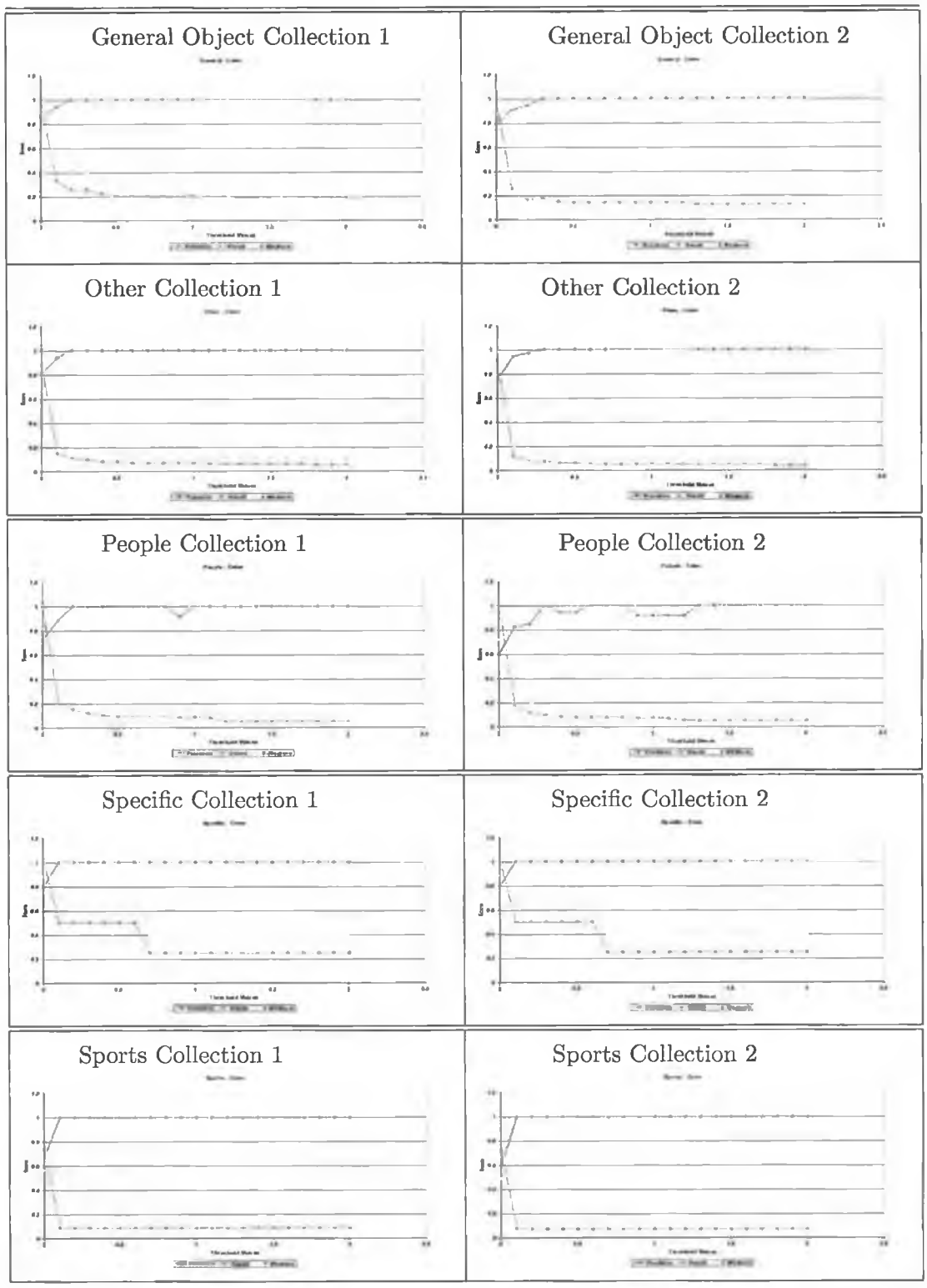

Figure D.16: Threshold variation graphs over the both Collection_1 and Collection_2 for low level features "HSVColor" run

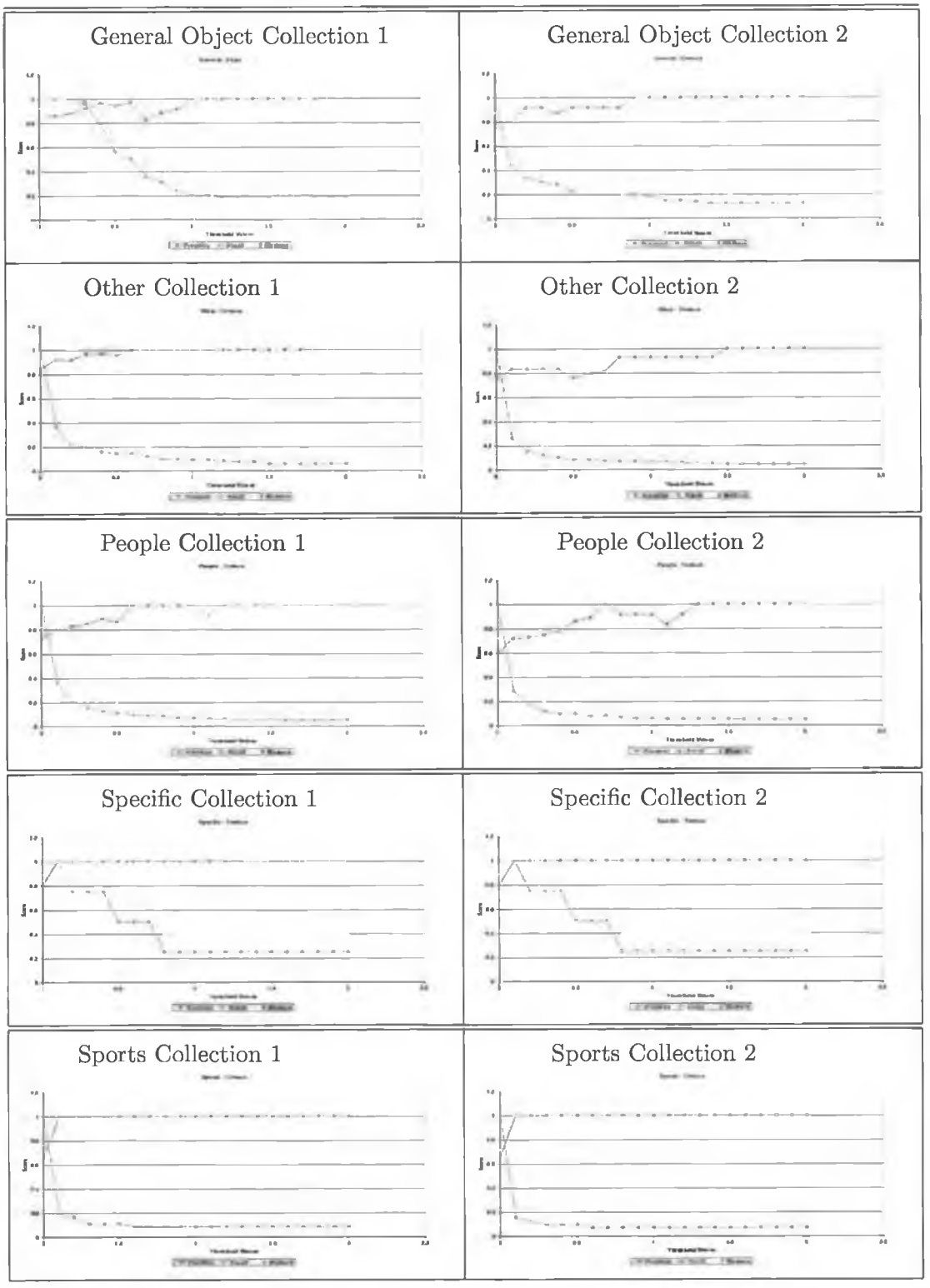

Figure D.17: Threshold variation graphs over the both Collection_1 and Collection_2 for low level features "Texture" run

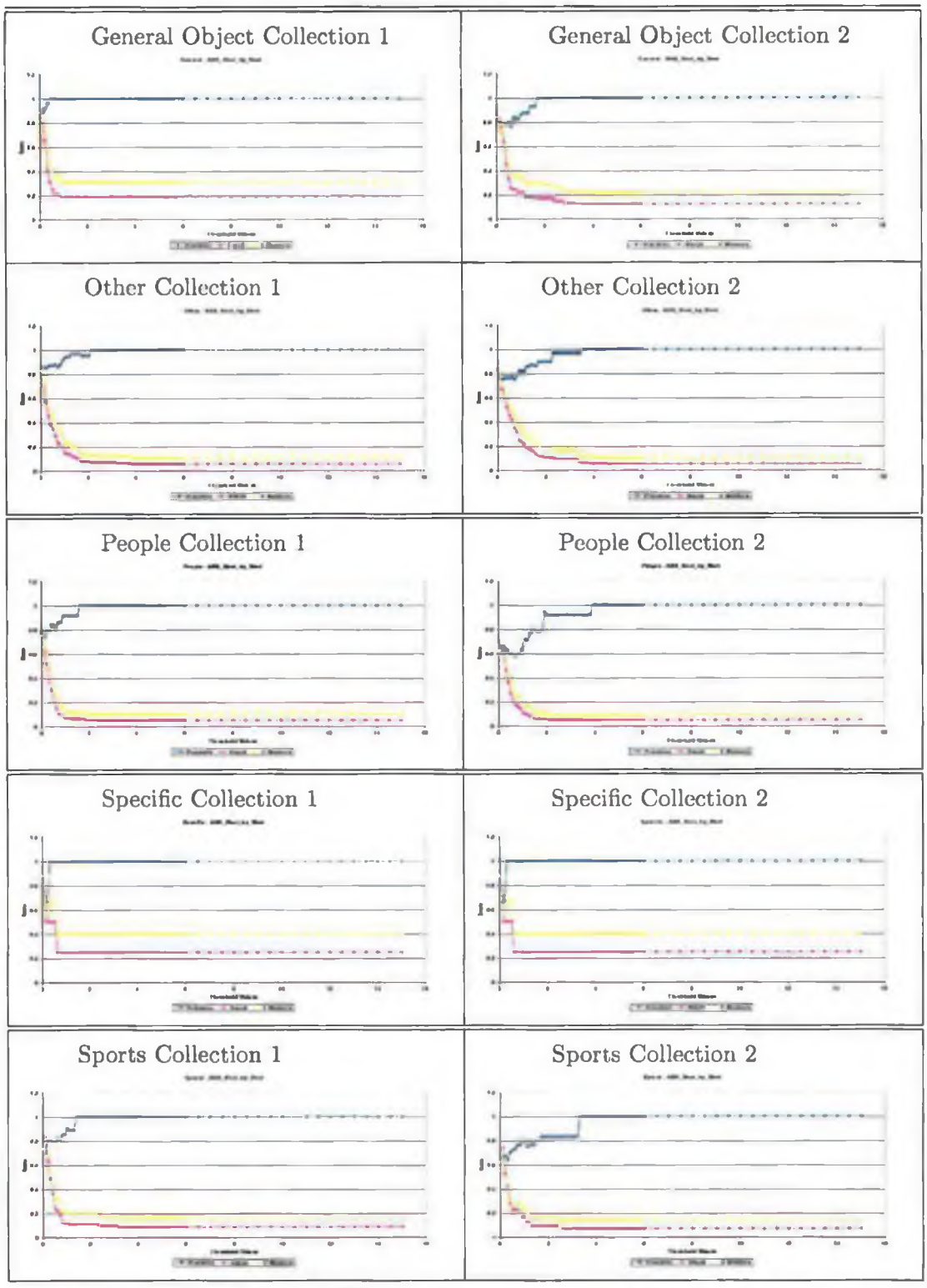

Figure D.18: Threshold variation graphs over the both Collection_1 and Collection_2 for ASR transcript resources using a shot by shot approach to novelty detection "ASR_Shot_by_Shot" run

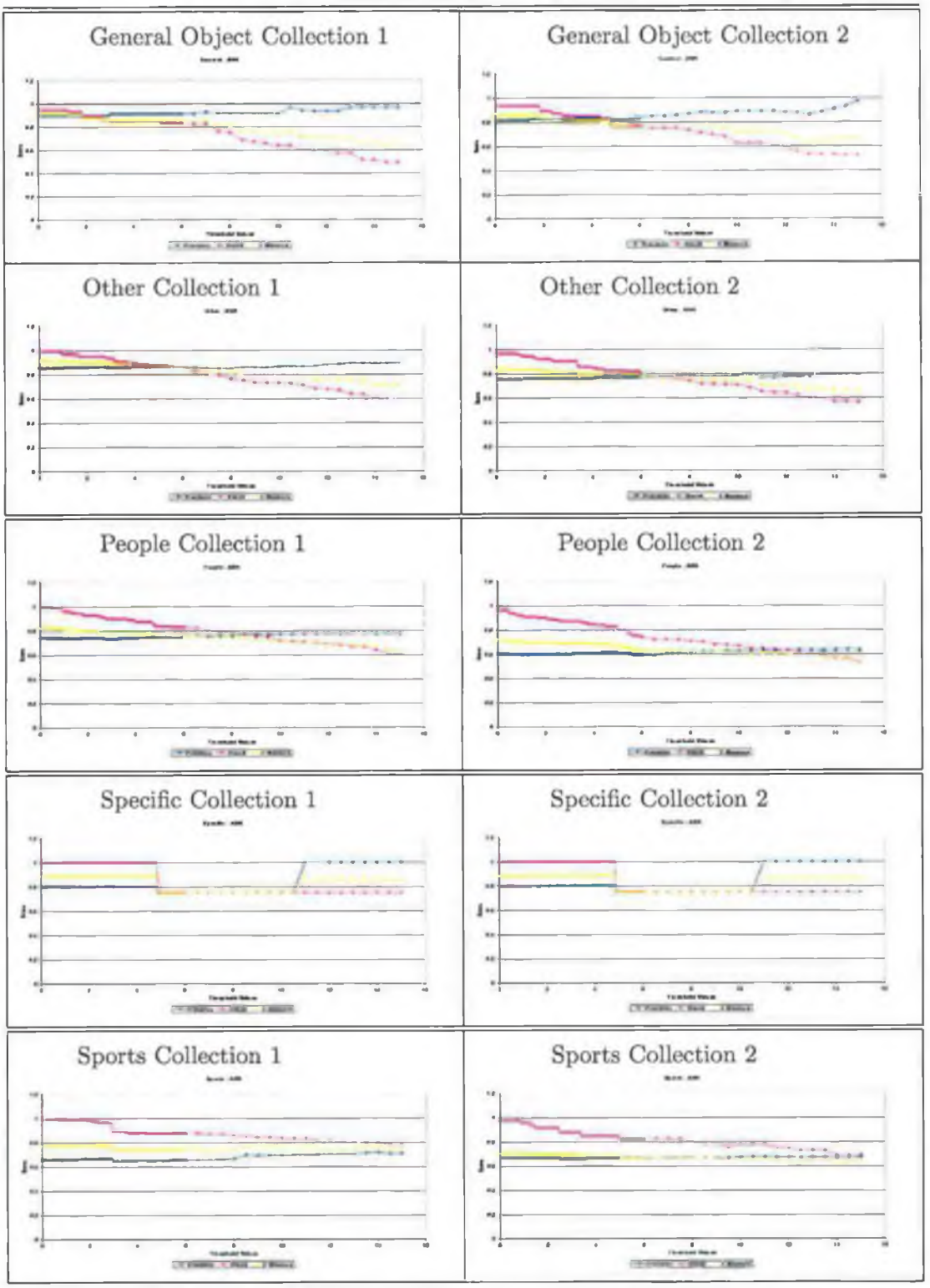

Figure D.19: Threshold variation graphs over the both Collection_1 and Collection_2 for ASR transcript resources using an accumulative history approach to novelty detection "ASR" run

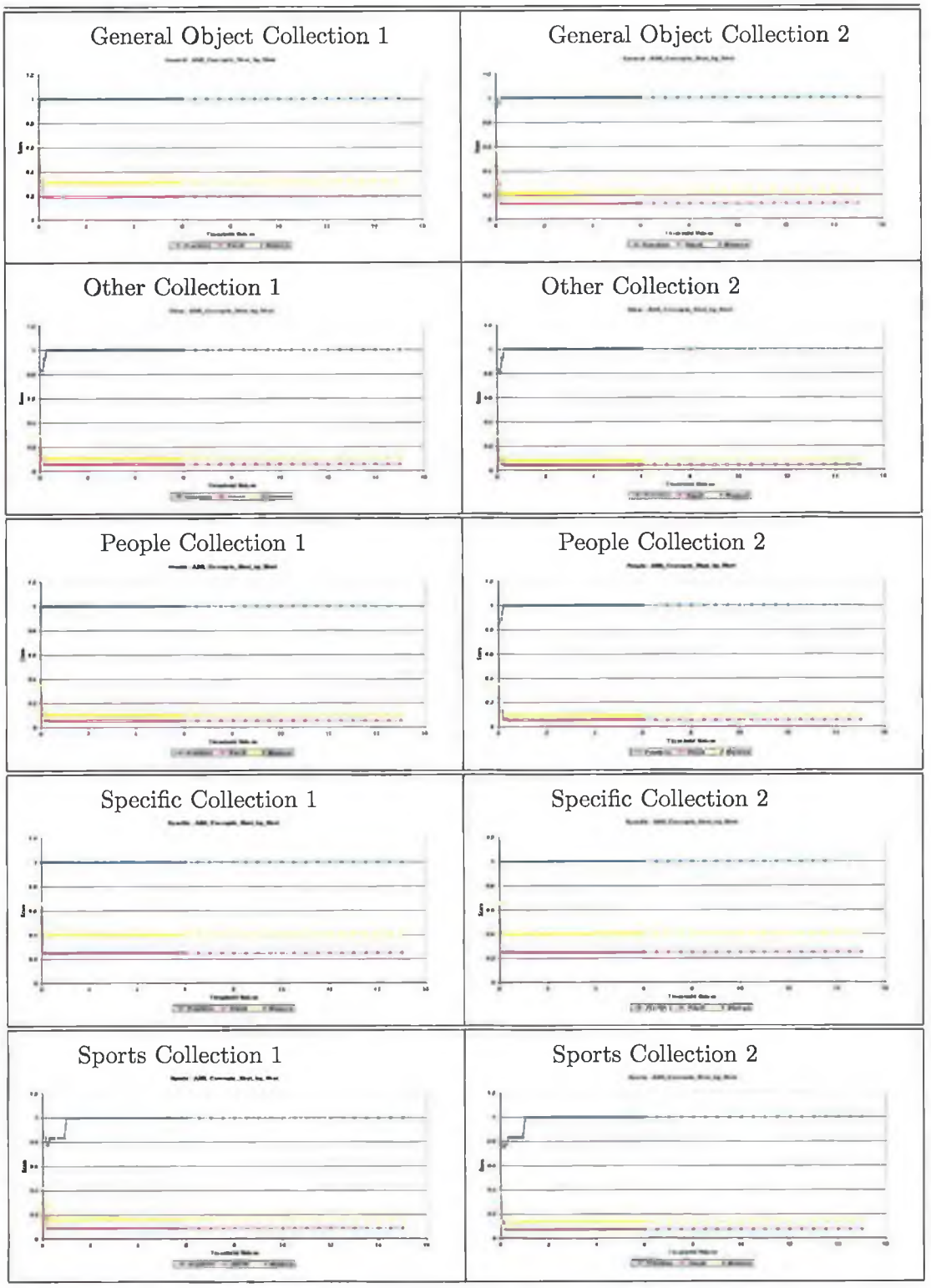

Figure D.20: Threshold variation graphs over the both Collection_1 and Collection_2 for ASR transcript and manual concept resources using a shot by shot approach to novelty detection "ASR_Concepts_Shot_by_Shot" run

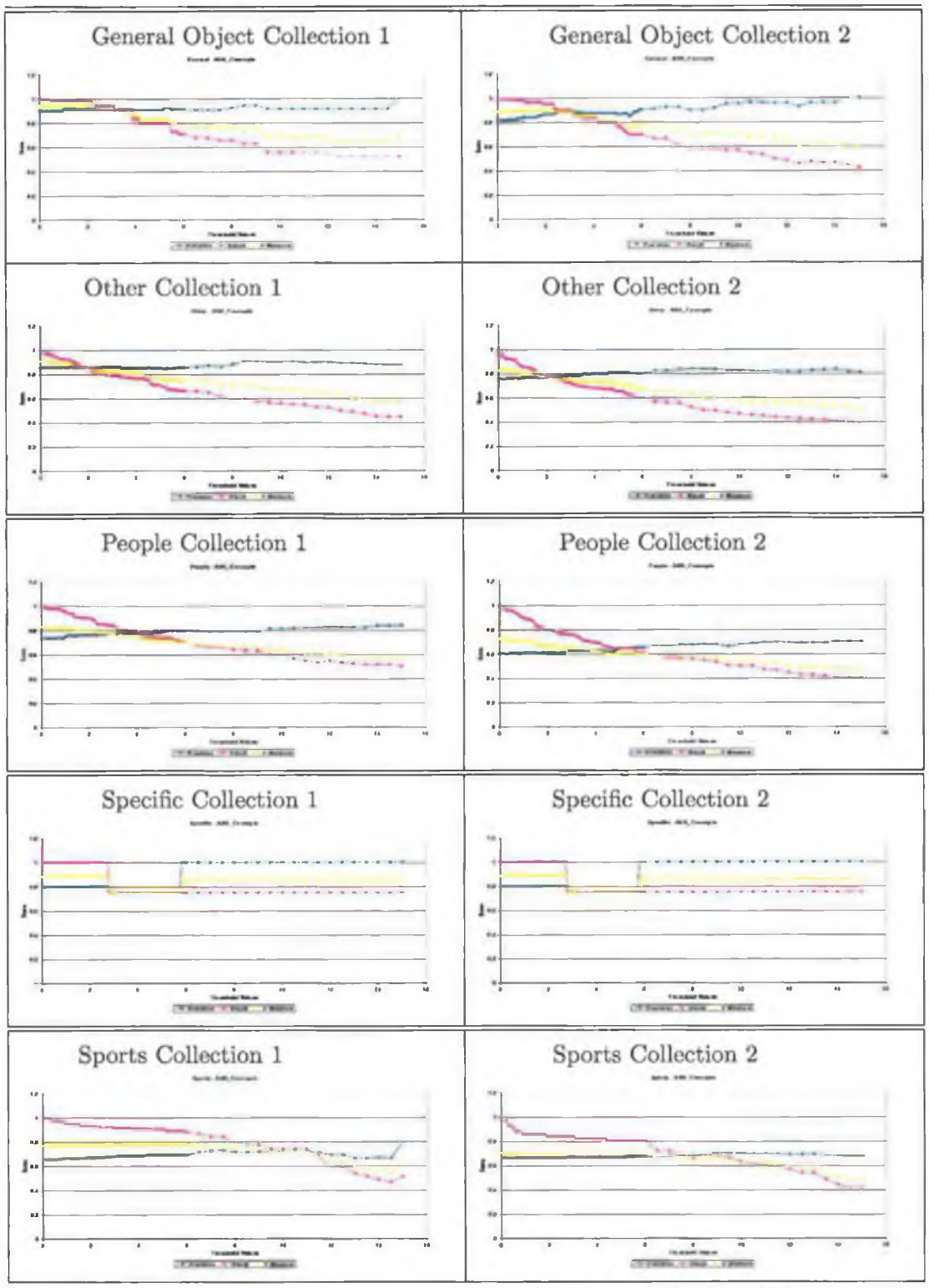

Figure D.21: Threshold variation graphs over the both Collection_1 and Collection_2 for ASR transcript and manual concept resources using an accumulative history approach to novelty detection "ASR_Concepts" run

Appendix E

Experimental Run Median difference Graphs

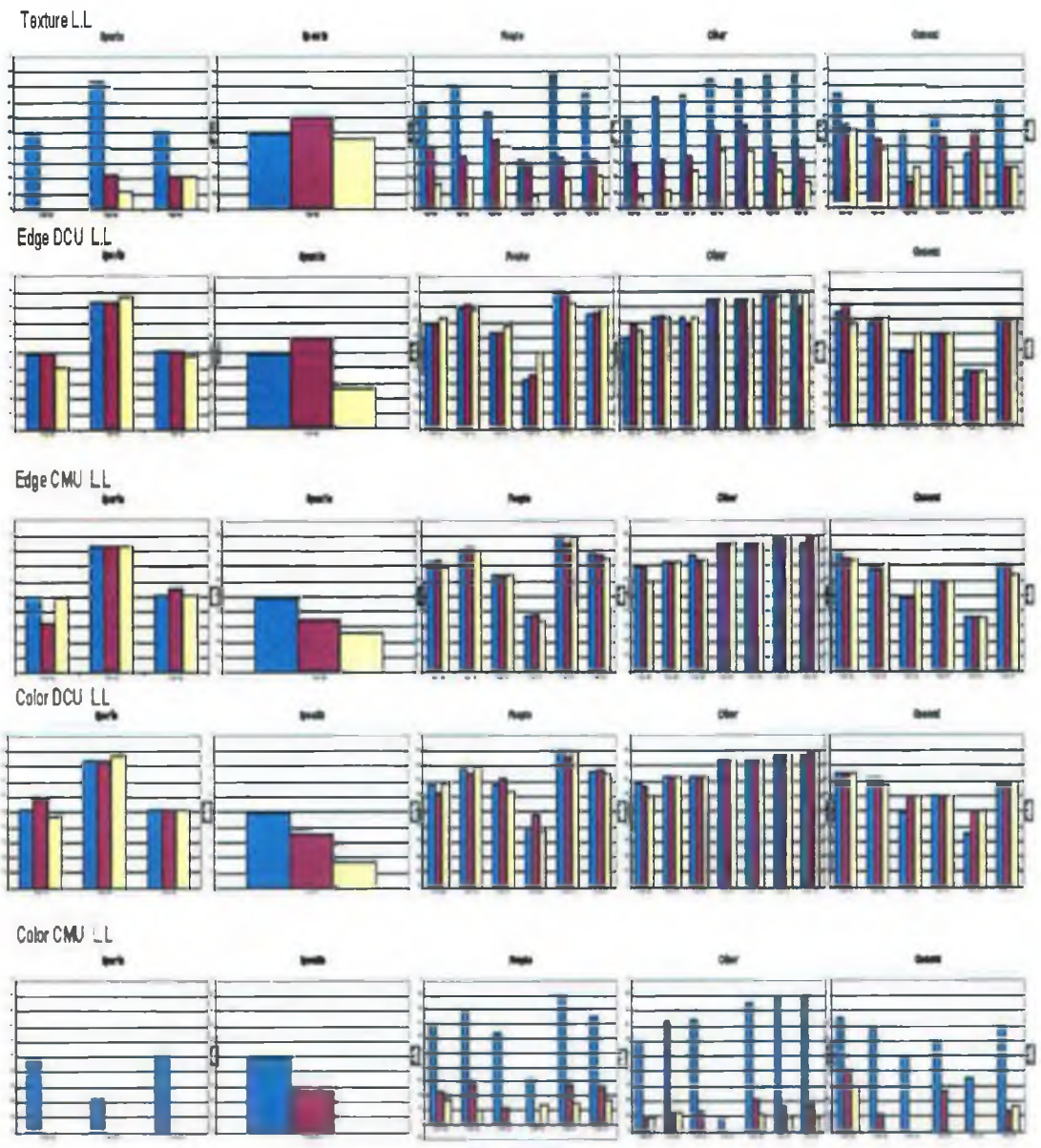

Figure E.1: Median Difference graphs of low level feature runs over Collection_1

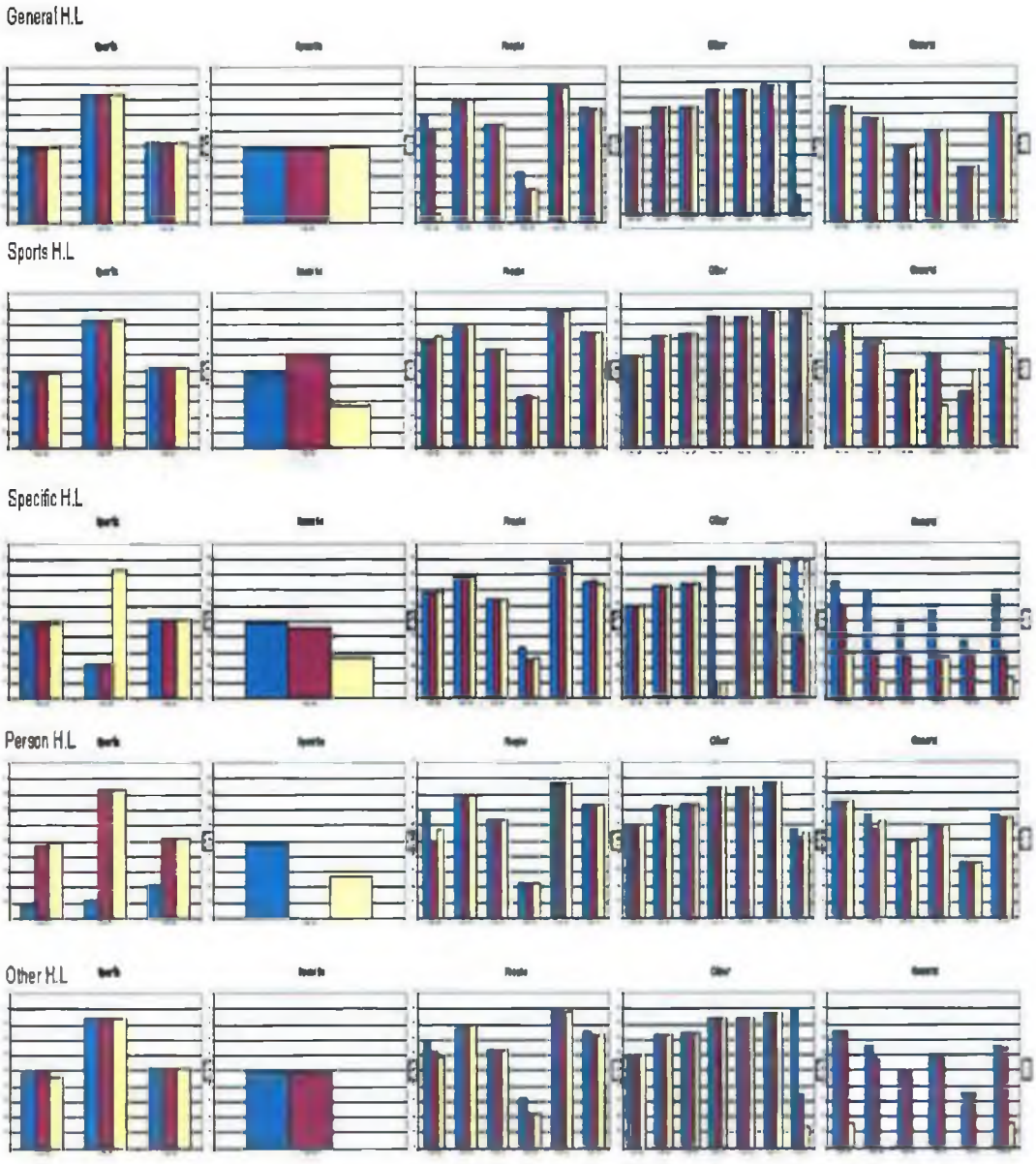

Figure E.2: Median Difference graphs of high level feature runs over Collection_1

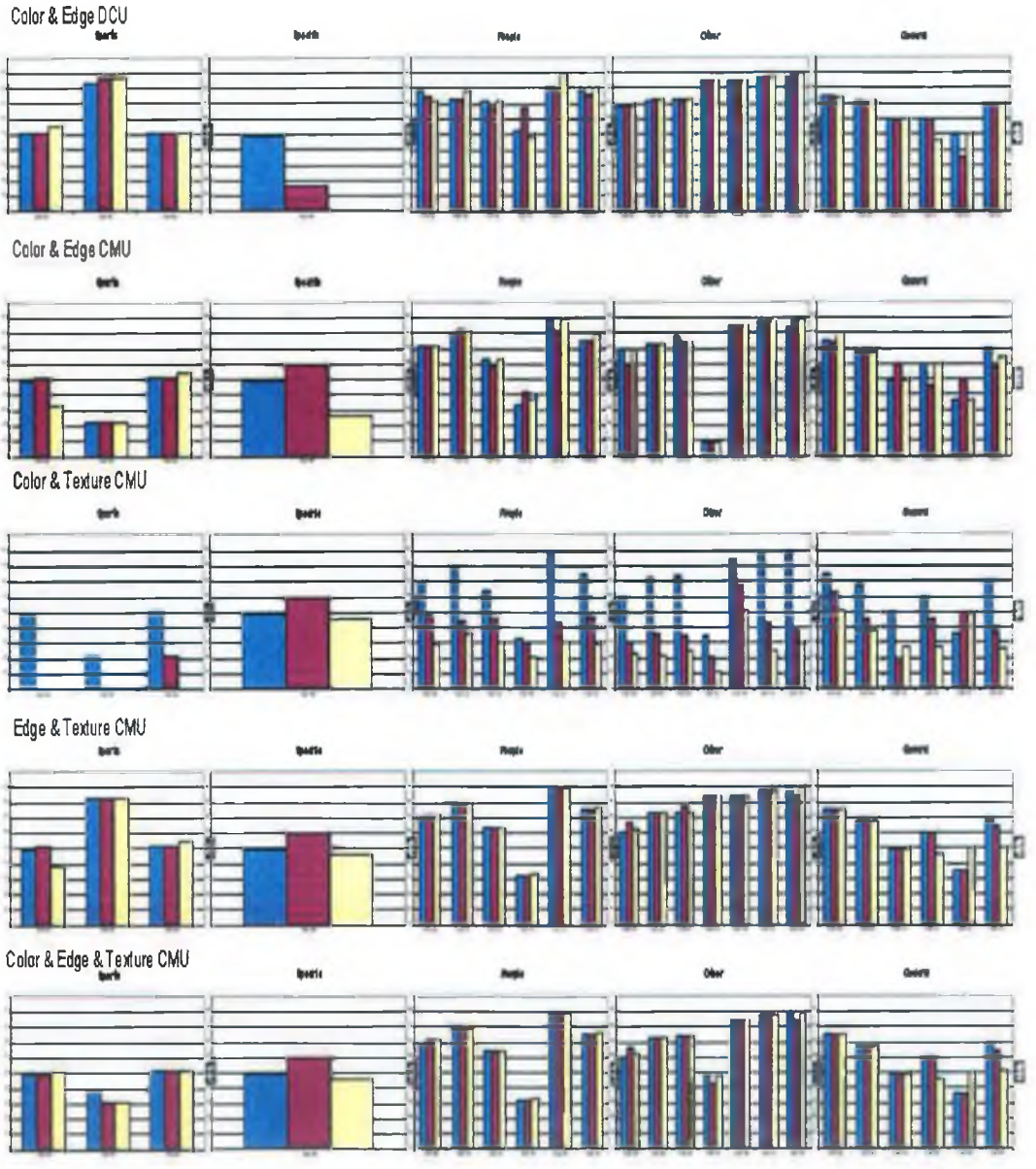

Figure E.3: Median Difference graphs of low level combination runs over Collection_1

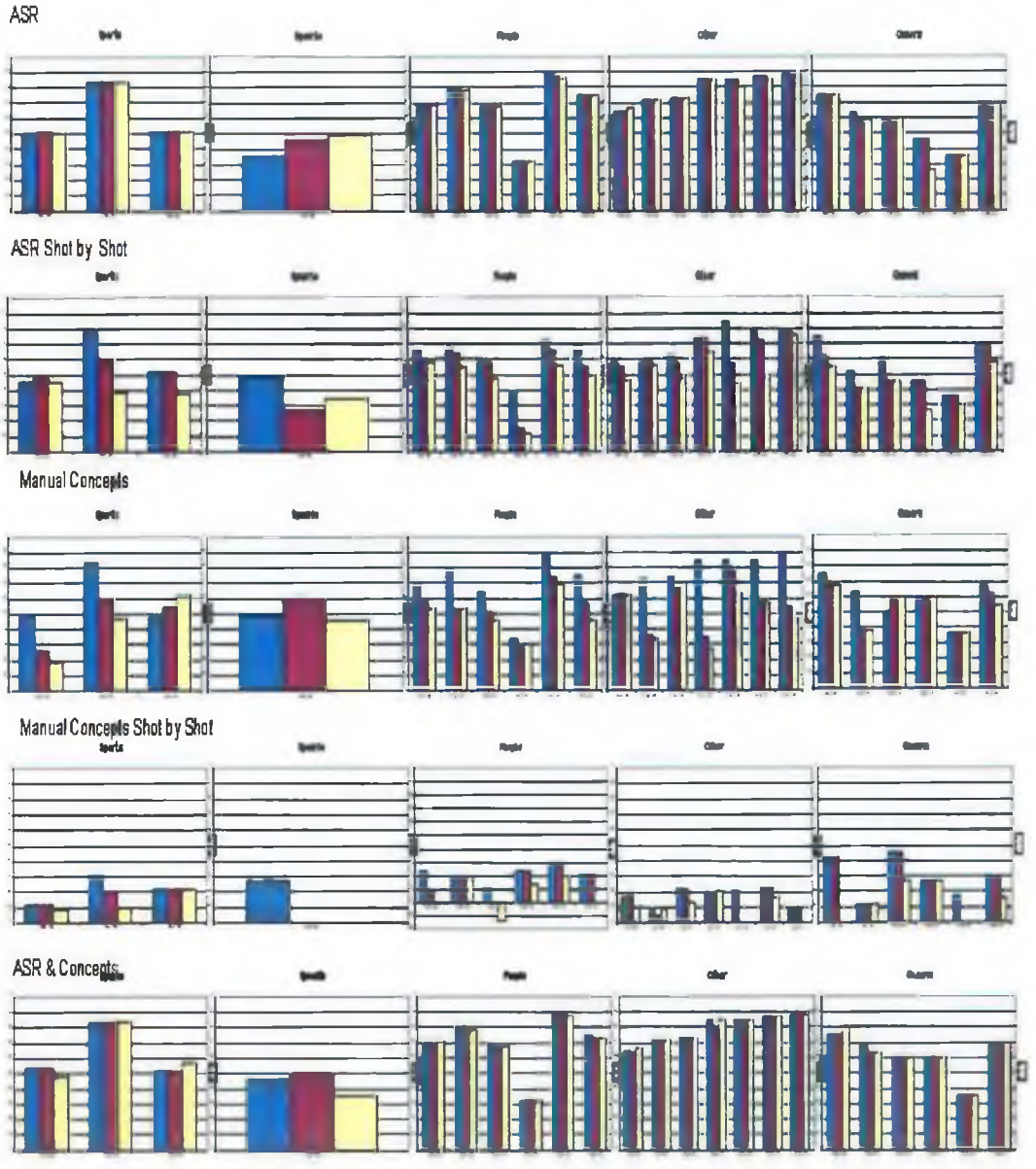

Figure E.4: Median Difference graphs of ASR and manually annotated runs over Collection_1

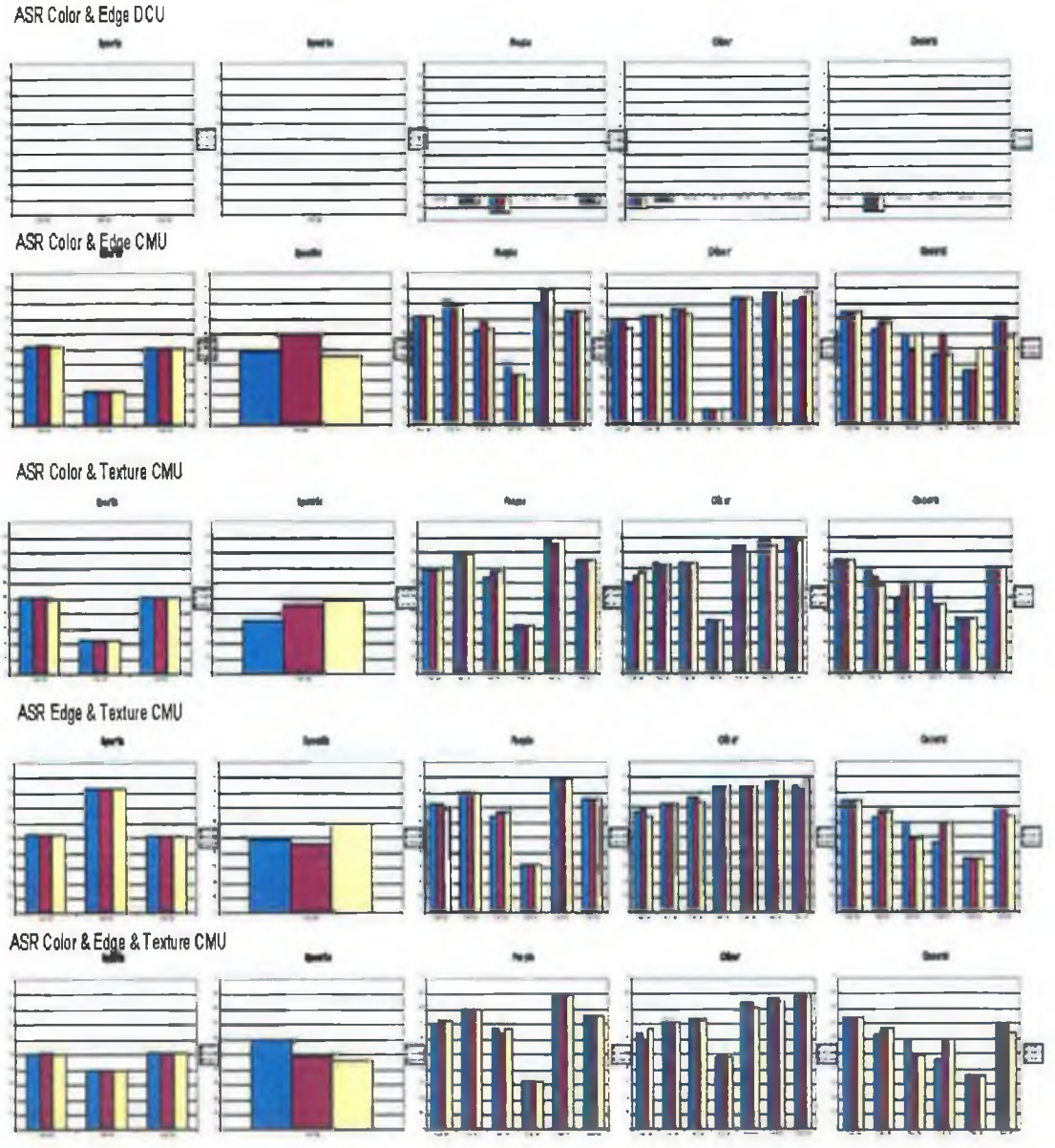

Figure E.5: Median Difference graphs of ASR low level combination runs over Collection_1

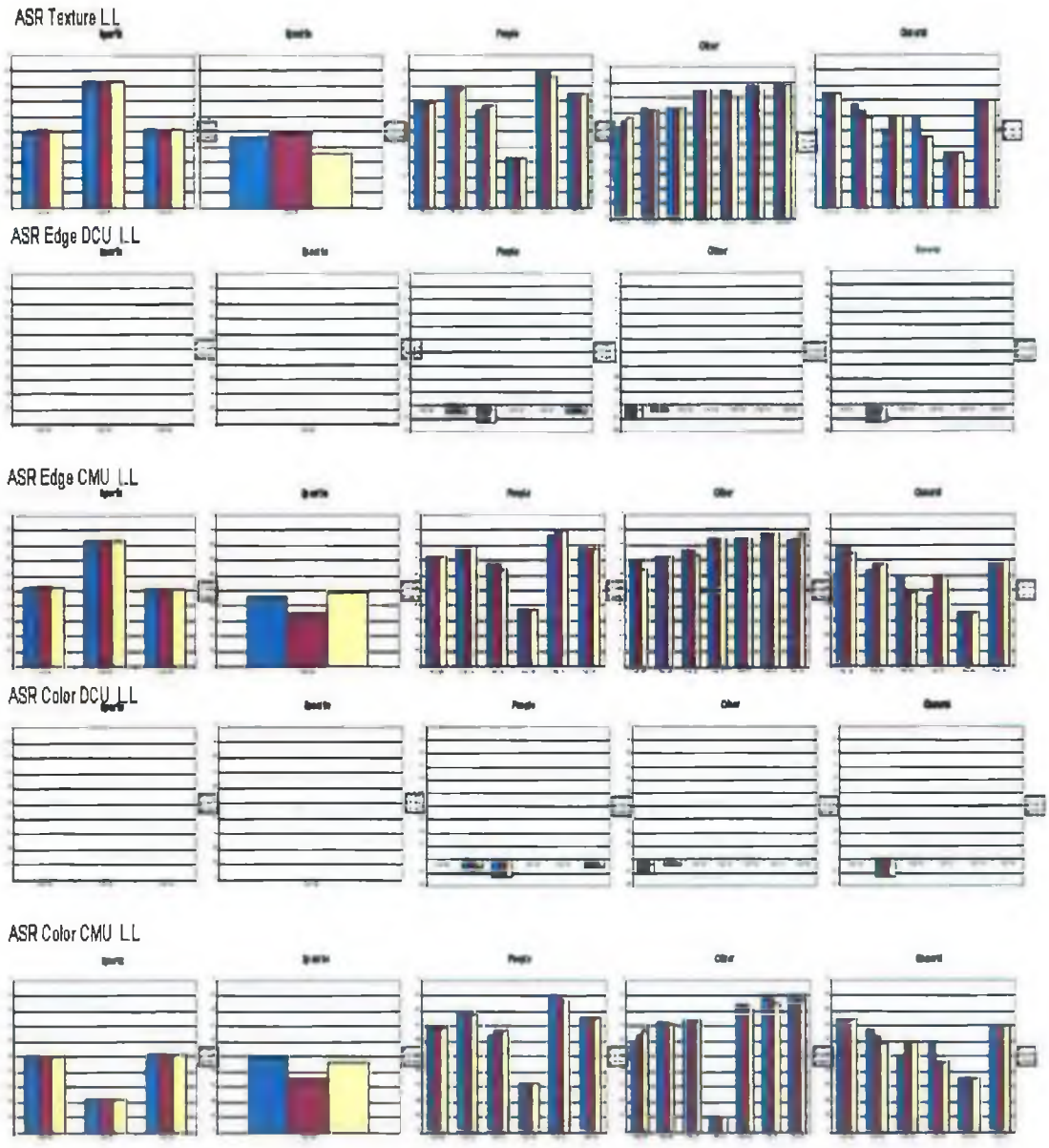

Figure E.6: Median Difference graphs of ASR low level runs over Collection_1

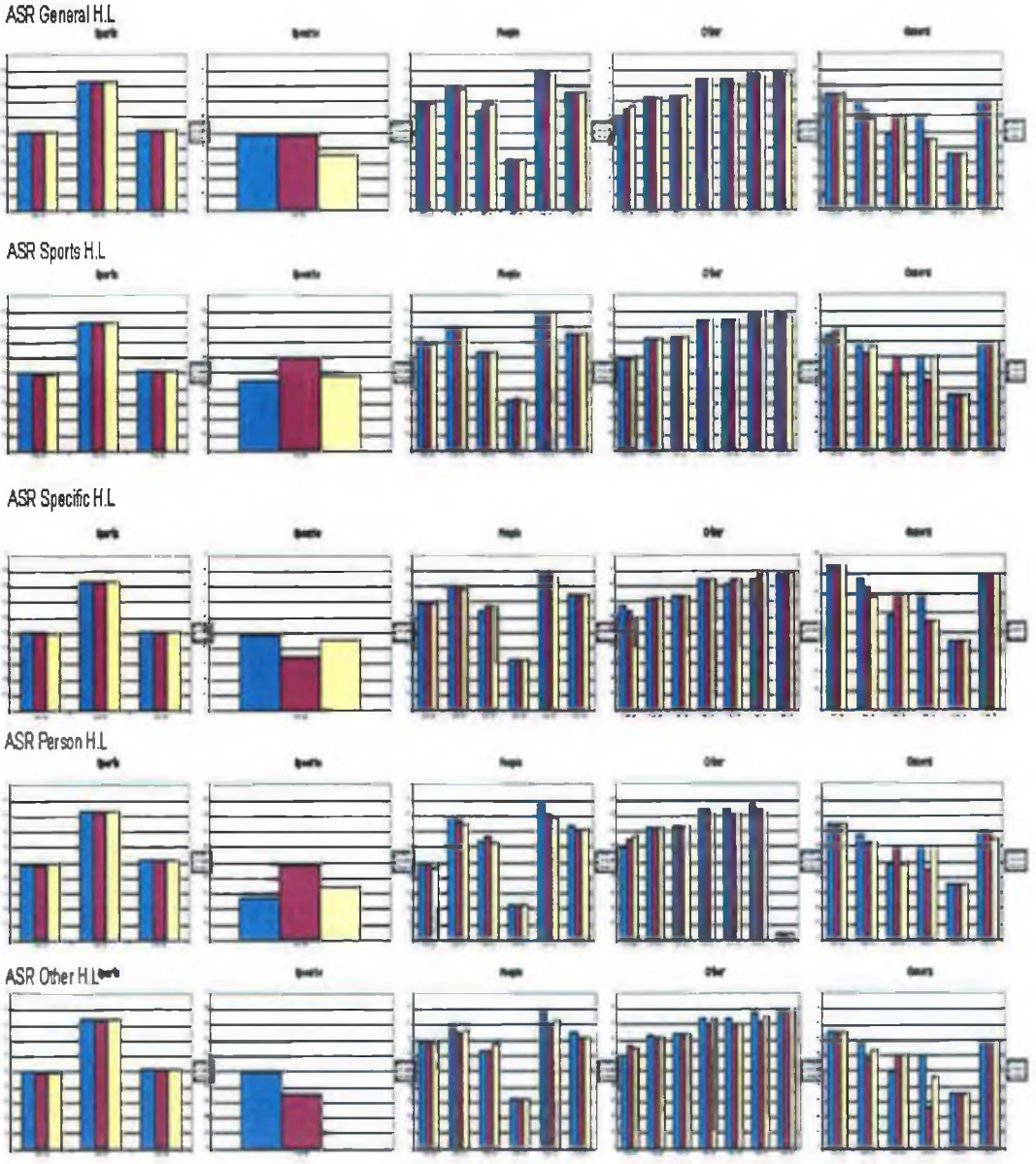

Figure E.7: Median Difference graphs of ASR high level runs over Collection_1

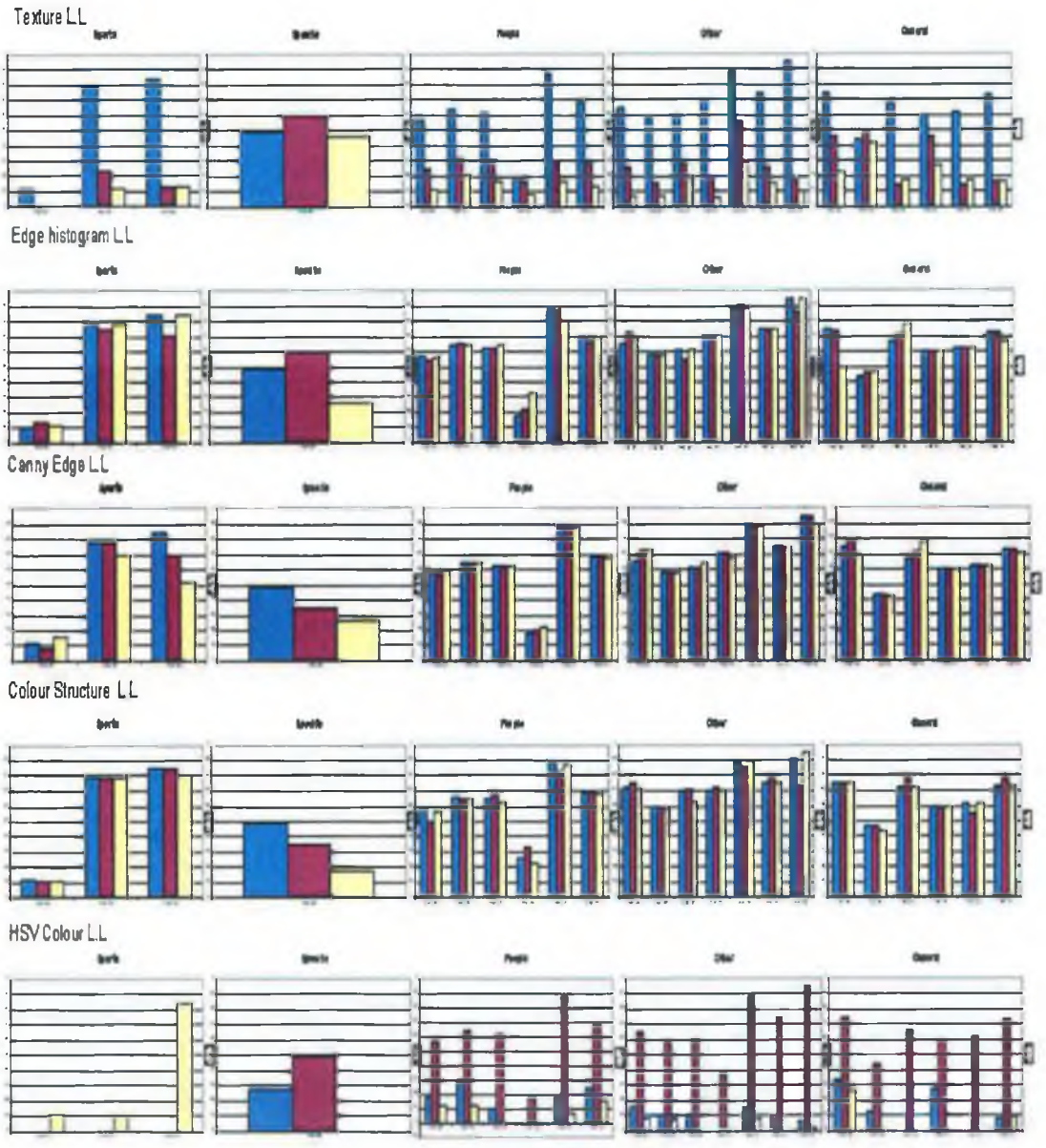

Figure E.8: Median Difference graphs of low level feature runs over Collection_2

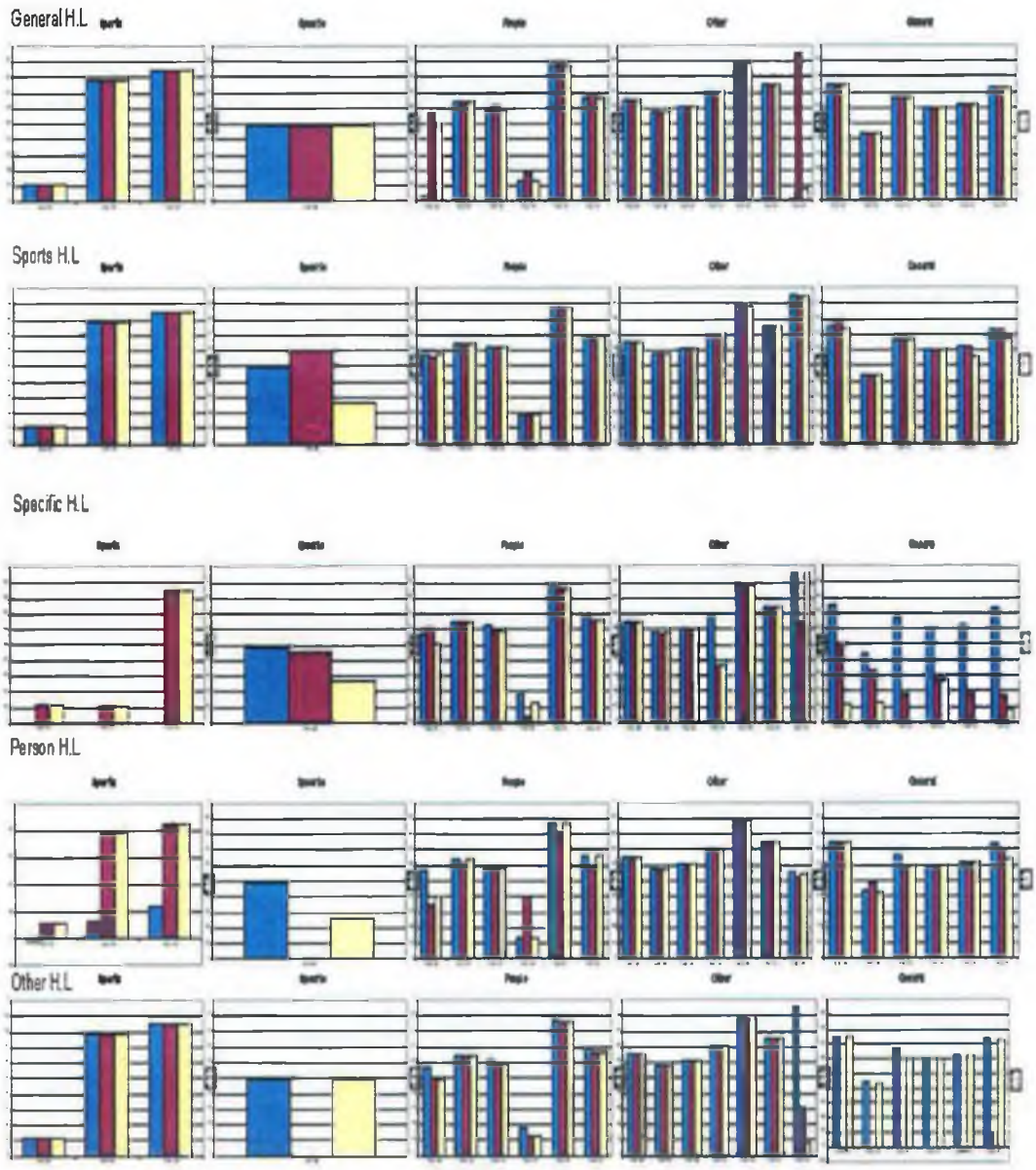

Figure E.9: Median Difference graphs of high level feature runs over Collection_2

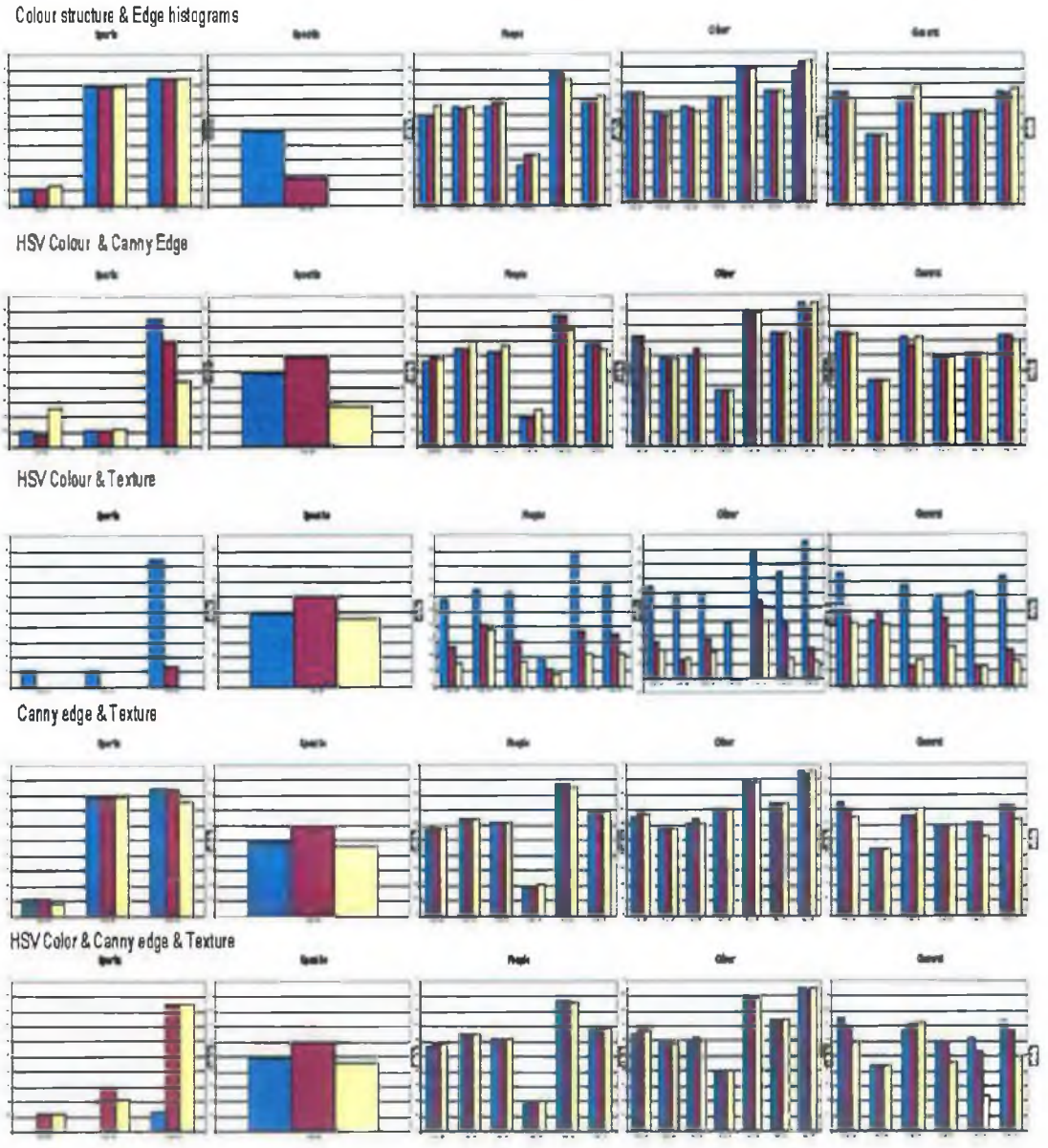

Figure E.10: Median Difference graphs of low level combination runs over Collection_2

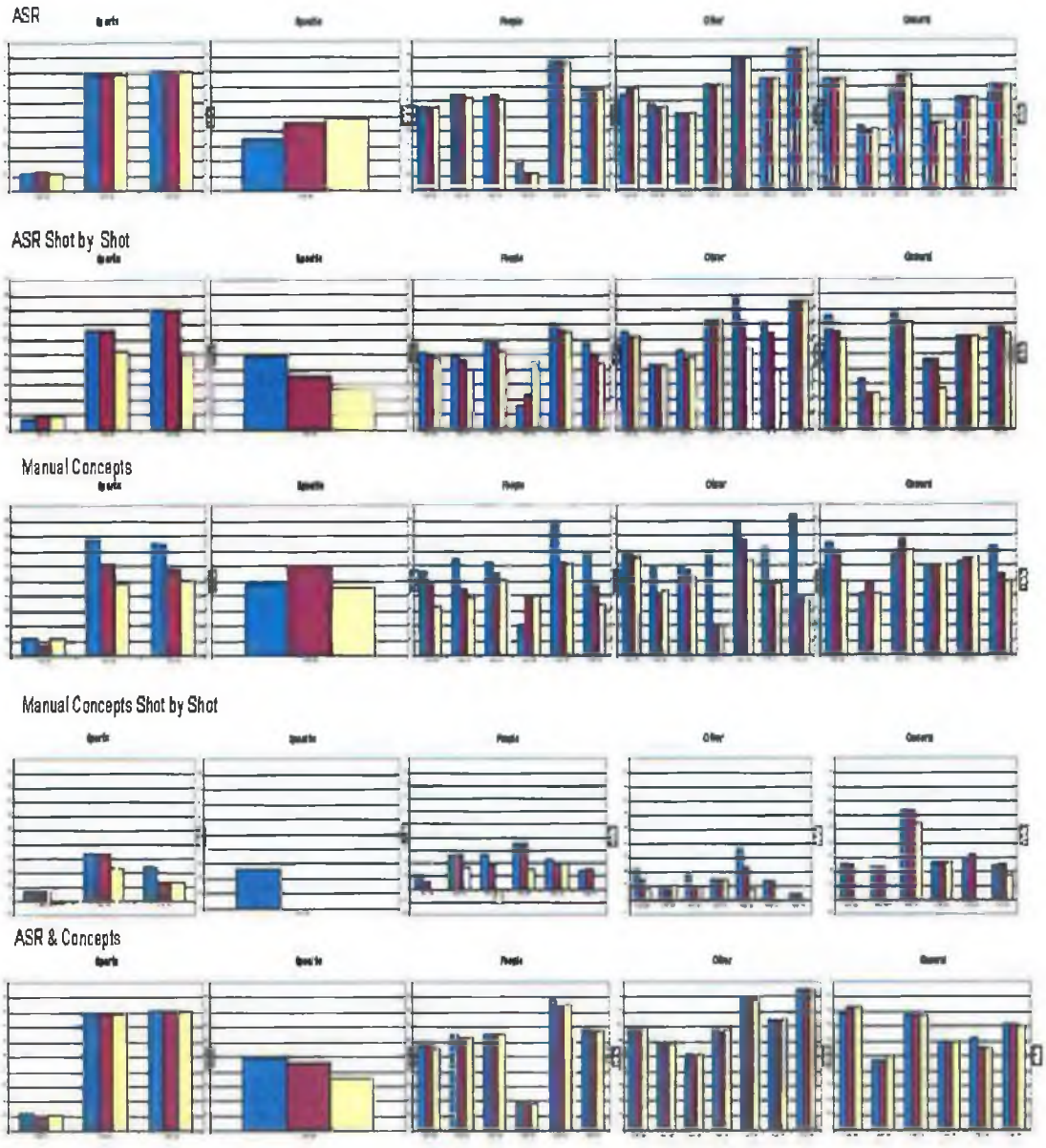

Figure E.11: Median Difference graphs of ASR and manually annotated runs over Collection_2

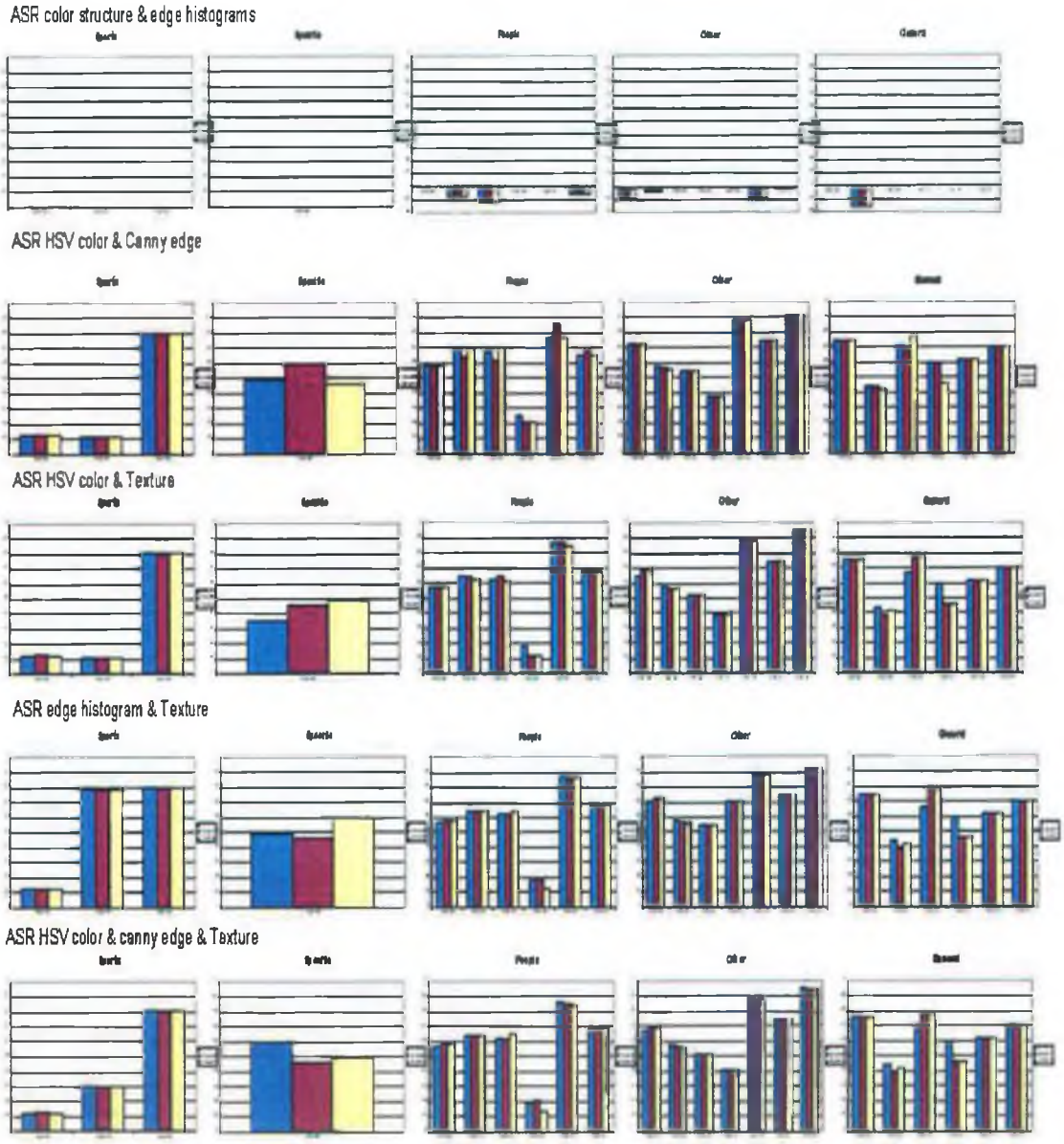

Figure E.12: Median Difference graphs of ASR low level combination runs over Collection_2

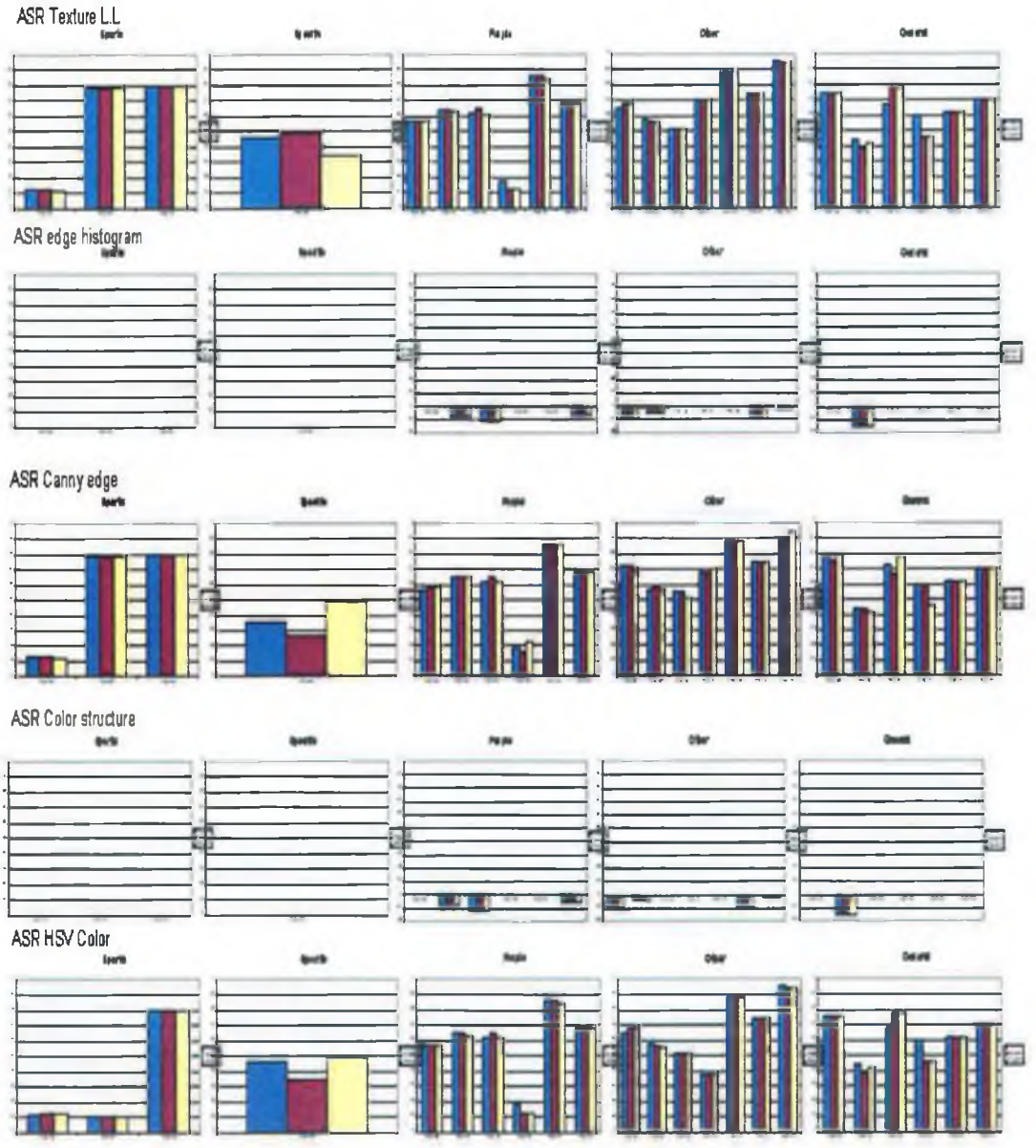

Figure E.13: Median Difference graphs of ASR low level runs over Collection_2

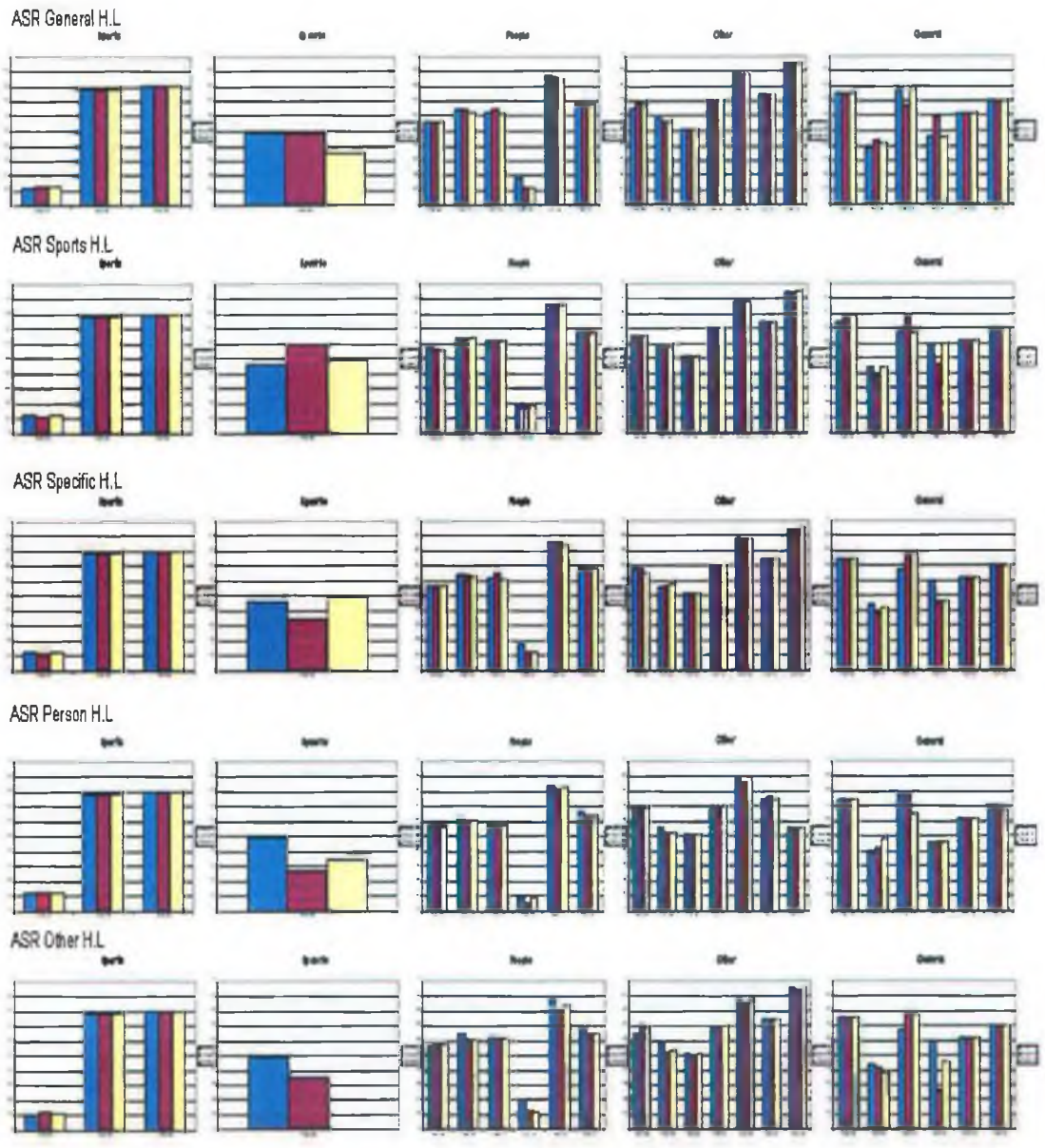

Figure E.14: Median Difference graphs of ASR high level runs over Collection_2

References

[AAB+02] J. Allan, J. Aslam, N. Belkin, C. Buckley, J. Callan, B. Croft, S. Dumais, N. Fuhr, D. Harper, D. Hiemstra, W. Kraail, D. Harman, E. Hovy, D. Lewis, T . Hofmannand J. Lafferty, V. Lavrenko, L. Liddy, A. McCallum, R. Manmatha, J. Ponte, J. Prager, D. Radev, P. Resnik, S. Robertson, R. Rosenfeld, S. Roukos, M. Sanderson, R. Schwartz, A. Singhal, A. Smeaton, H. Turtle, R. Weischedel, E. Voorhees, J. Xu, and C. Zhai. Challenges in information retrieval and language modeling. In *Report of Workshop held Center for Intelligent Information Retrieval, University of Massachusetts Amherst*, 2002.

[AE96] L. H. Armitage and P. G. B. Enser. Information need in the visual document domain:. In *Report on Project RDD/G/235 to the British Library Research and Innovation Centre*, 1996.

[AGK01] James Allan, Rahul Gupta, and Vikas Khandelwal. Temporal summaries of new topics. In *SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 10–18, New York, NY, USA, 2001. ACM Press.

[Ago02] M Agosti. Information retrieval on the web. In *Proceedings of the European Summer School in Information Retrieval 2000 , Lectures on Information Retrieval*, pages 242–285. Springer, 2002.

[AJAC+04] N. Abdul-Jaleel, J. Allan, W. B. Croft, F. Diaz, L. Larkey, X. Li, D. Metzler, M. D. Smucker, T. Strohman, H. Turtle, and C. Wade. UMass at TREC 2004: Notebook. In *Proceedings of the 13th Text Retrieval Conference (TREC 2004)*, 2004.

[AJR+99] James Allan, H. Jin, M. Rajman, C. Wayne, D. Gildea, V. Larenko, R. Hoberman, and D. Caputo. Topic-based novelty detection. In *Final Report of Summer workshop at center for language and speech processing St. John Hopkins University*, 1999.

[Arc02] The Internet Archive. Movie archive homepage. In *URL: http://www.archive.org/movies/*, 2002.

[AWB03] James Allan, Courtney Wade, and Alvaro Bolivar. Retrieval and novelty detection at the sentence level. In *SIGIR 03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 314–321, New York, NY, USA, 2003. ACM Press.

[BCG+03] P. Browne, C. Czirjek, G. Gaughan, C. Gurrin, G. Jones, H. Lee, S. Marlow, K. McDonald, N. Murphy, N. O'Connor, N. O'Hare, A. F. Smeaton, and J. Ye. Dublin City University video track experiments for TREC 2003. In *TRECVID 2003: - Text REtrieval Conference TRECVID Workshop*, 2003.

[BCG+04] P. Browne, C. Czirjek, G. Gaughan, C. Gurrin, G. Jones, H. Lee, S. Marlow, K. McDonald, N. Murphy, N. O'Connor, N. O'Hare, A. F. Smeaton, and J. Ye. Dublin City University video track experiments for TREC 2004. In *TRECVID 2004: - Text REtrieval Conference TRECVID Workshop*, 2004.

[Bim99] A. Del Bimbo. *Visual Information Retrieval.* Academic Press, 1999.

[BMT93] R. Beckwith, G. A. Miller, and R. Tengi. Design and implememtation of the wordnet lexical database and searching software. In *Five Papers on WordNet*, pages 62–77, 1993.

[Can86] J. Canny. A computational appraoch to edge detection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 8,No. 6*, 1986.

[CDV04] CDVP2004. www.cdvp.dcu.ie. In *Center for Digital Video Processing Homepage*, 2004.

[CG98] Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR 98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, New York, NY, USA, 1998. ACM Press.

[Com02] MPEG-7 Committee(2002). MPEG-7: Multimedia content description interface. In *ISO*, 2002.

[Dom01] S. Dominich. *Mathematical Foundations of Information Retrieval.* Kluwer Academic Publishers, 2001.

[DT] Topic Detection and Tracking Homepage TDT. Url: http://www.nist.gov/speech/tests/tdt/.

[Ear85] R. A. Earnshaw. *Fundamental Algorithms for Computer Graphics. Chapter on Visual Perception and Computer Graphics page 1006.* Springer-Verlag, 1985.

[Ens95] P. G. B. Enser. Pictorial information retrieval. In *Journal of Documentation 51(2) pp. 126-170*, 1995.

[Gan80] H.J. Gans. *Deciding Whats News: A Case study of CBS Evening News, NBC Nightly News, Newsweek and Time.* Vintage Books, 1980.

[GSG+03] Georgina Gaughan, Alan F. Smeaton, Cathal Gurrin, Hyowon Lee, and Kieran McDonald. Design, implementation and testing of an interactive video retrieval system. In *MIR'03: Proceedings of the ACM Multimedia Information retrieval workshop*, New York, NY, USA, 2003. ACM Press.

[gui] TrecVid 2005 guidelines. http://www-nlpir.nist.gov/projects/tv2005/tv2005.html.

[GW92] R. Gonzalez and R. Woods. Digital image processing. In *Addison Wesley pp. 414-428*, 1992.

[Har92] D. Harman. The DARPA TIPSTER project. In *SIGIR '92: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 26–28, 1992.

[Har02] D. Harman. Overview of TREC 2002 novelty track. In *Proceedings of the 11th Text Retrieval Conference TREC 2002*, 2002.

[Hau04] A. G. Hauptmann. Towards a large scale concept ontology for broadcast video. In *Proceedings of the International Conference on Image and Video Retrieval CIVR04*, pages 674–675, 2004.

[Hie01] D. Hiemstra. Using language models for information retrieval. In *PhD Thesis*, University of Twente, The Netherlands, 2001.

[Hom05] The TRECVid Guidelines HomePage. Url: http://www-nlpir.nist.gov/projects/tv2005/tv2005.html. 2005.

[JGA02] L. Lamel J.L. Gauvain and G. Adda. The LIMSI broadcast news transcription system. In *Speech Communication, 37(1-2)*, pages 89–108, 2002.

[JS03] A. Jaimes and J. R. Smith. Semi-automatic, data-driven construction of multimedia ontologies. In *Proceedings of the 2003 IEEE International Conference on Multimedia and Expo ICME*, 2003.

[JZX03] Q. Jin, J. Zhao, and B. Xu. Nlpr at TREC 2003: Novelty and robust. In *Proceedings of the 12th Text Retrieval Conference (TREC 2003)*, 2003.

[KKK+04] K. Karoji, T. Kondo, Y. Kakuta, T. Tomiyama, and T. Takagi. Meiji University: Web, novelty and genomics track experiments. In *Proceedings of the 13th Text Retrieval Conference (TREC 2004)*, 2004.

[KN04] J. R. Kender and M. R. Naphade. Ontology design for video semantic threads. In *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo ICME*, 2004.

[KSG06] M. Koskela, A.F. Smeaton, and G. Gaughan. Semantic analysis of concept models for news videos. In *Proceedings of Categorisation and Mulitmedia knowledge management systems*, 2006.

[Lay94] Sara Shatford Layne. Some issues in the indexing of images. In *Journal of the American Society for Information Science 45(8) pp. 583-588*, 1994.

[LSM+01] H. Lee, A. F. Smeaton, N. Murphy, S. Marlow, and N. O'Connor. User interface design for keyframe-based browsing of digital video. In *WIAMIS 2001: Workshop on Image Analysis for Multimedia Interactive Services*, 2001.

[LSO+00] H. Lee, A. F. Smeaton, C. O'Toole, N. Murphy, S. Marlow, and N. O'Connor. The Físchlár digital video recording, analysis, and browsing system. In *RIAO 2000: Content-based Multimedia Information Access, Paris, France,*, pages 26–28, 2000.

[LTS02] C. Y. Lin, B. L. Tseng, and J. R. Smith. IBM MPEG-7 annotation tool. In *Proceedings of the 2003 IEEE International Conference on Multimedia and Expo ICME*, 2002.

[LTS03] C. Y. Lin, B. L. Tseng, and J. R. Smith. Video collaborative annotation forum: Establishing ground -truth lables on large multimedia datasets. In *Proceedings of NIST TRECVID2003 Video Retrieval Conference TRECVID 2003*, 2003.

[Mar01] G. Marchionini. The open video project homepage. In *URL: http://www.open-video.org*, 2001.

[MS05] K. McDonald and A. F. Smeaton. A comparison of score, rank and probability-based fusion methods for video shot retrieval. In *Proceedings of the International Conference on Image and Video Retrieval CIVR05*, 2005.

[Nis] NIST Digital Video Collection Vol-1. In *URL: http://www.nist.gov/srd/nistsd26.htm.*

[NS04] M. R. Naphade and J. R. Smith. On the detection of semantic concepts at TRECVID. In *Proceedings of 12th ACM international conference on Multimedia*, pages 660–667. ACM Press, 2004.

[OMM+01] N. O'Connor, S. Marlow, N. Murphy, A. F. Smeaton, P. Browne, S. Deasy, H. Lee, and K. Mc Donald. Físchlár: An on-line system for indexing and browsing of broadcast television content. In *ICASSP 2001: International Conference on Acoustics, Speech, and Signal Procesing*, 2001.

[OPH96] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based feature distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[pag] MPEG Home page. Url: http://www.chiariglione.org/mpeg/.

[PHB97] J. Puzicha, T. Hofmann, and J. Buhmann. Non-parametric similarity measures for unsupervised texture segmentation and image retrieval. pages 267–272, 1997.

[PHO+02] M. Pickering, D. Heesh, R. O'Callaghan, S. Ruger, and D. Bull. Video retrieval using global features in keyframes. In *In Proceedings of the Eleventh Text REtrieval Conference(TREC2002)*, 2002.

[Poy] Charles Poynton. Colour space information and conversions. Technical report.

[Rao82] C. Radhakrishna Rao. Diversity: Its measurement, decomposition, apportionment and analysis. *Sankyha: The Indian Journal of Statistics*, 44(A):1–22, 1982.

[RHH+04] M. Rautiainen, M. Hosio, I. Hanski, M. Varanka, J. Kortelainen, T. Ojala, and T. Seppanen. TRECVID2004 experiments at team oula. In *TRECVID 2004, NIST publications*, 2004.

[RM84] J. Ratcliff and D. Metzener. Pattern matching: The Gestalt approach. In *Dr. Dobbs Journal*, page 46, 1984.

[RNBY99] Ribeiro-Neto. and R. Baeza-Yates. *Modern Information Retrieval.* Addison-Wesley, 1999.

[Rob77] S. E. Robertson. The probability ranking principal in ir. In *Journal Documentation 33(4)*, pages 294–304, 1977.

[RTE02] RTE. Annotation guidelines for broadcast tv news. In *Personal Communication*, 2002.

[RWB+97] S. E. Robertson, S. Walker, M. Boughanem, G. Jones, and K. Sparck Jones. Okapi at TREC-6 automatic ad hoc, vlc, routing, filtering and qsdr. In *Proceeding of the 6th Annual TREC Conference*, 1997.

[Sal89] G. Salton. *Automatic Test Processing: The Transformation, Analysis and Retrieval of Information by Computer.* Addison-Wesley, 1989.

[SB91] M. J. Swain and D. H. Ballard. Colour indexing. 7(1):11–32, 1991.

[SC01] N. Stokes and J. Carthy. Combining semantic and syntactic document classifiers to improve first story detection. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information and Retrieval*, 2001.

[SGL+04] A. F. Smeaton, C. Gurrin, H. Lee, K. Mc Donald, N. Murphy, N. O'Connor, D. O'Sullivan, B. Smyth, and D. Wilson. The Físchlár-News-Stories System: Personalised access to an archive of tv news. In *RIAO 2004: Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, 2004.

[SH03] Ian Soboroff and Donna Harman. Overview of the TREC2003 novelty track. In *Proceedings of the 12th Text Retrieval Conference (TREC 2003)*, 2003.

[SH04] Ian Soboroff and Donna Harman. Overview of the TREC2004 novelty track. In *Proceedings of the 13th Text Retrieval Conference (TREC 2004)*, 2004.

[Sha86] Sara Shatford. Analyzing the subject of a picture: a theoretical approach. In *Cataloging and Classification Quarterly, 6(3) pp. 39-62*, 1986.

[SKO03] A. F. Smeaton, W. Kraaij, and P. Over. TRECVID 2003 - an introduction. In *TRECVID 2003 - Text REtrieval Conference TRECVID Workshop*, 2003.

[SKO04a] A. F. Smeaton, W. Kraaij, and P. Over. The TREC video retrieval evaluation (TRECVID): A case study and status report. In *RIAO 2004 - Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, 2004.

[SKO04b] A. F. Smeaton, W. Kraaij, and P. Over. TRECVID 2004 - an overview. In *TRECVID 2004 - Text REtrieval Conference TRECVID Workshop*, 2004.

[SKO05] A. F. Smeaton, W. Kraaij, and P. Over. TRECVID 2005 - an overview. In *TRECVID 2005 - Text REtrieval Conference TRECVID Workshop*, 2005.

[Sme00] A. F. Smeaton. Indexing, browsing, and searching of digital video and digital audio information. In *European Summer School in Information Retrieval '2000 Lectures on Information Retrieval LNCS*, pages 93–110, 2000.

[SO02] A. F. Smeaton and P. Over. The TREC-2002 video track report. In *TRECVID 2002 - Text REtrieval Conference TRECVID Workshop*, 2002.

[SO03] A. F. Smeaton and P. Over. TRECVID: Benchmarking the effectiveness of information retrieval tasks on digital video. In *CIVR 2003 - International Conference on Image and Video Retrieval*, 2003.

[SOMM04] D. Sadlier, N. O'Connor, N. Murphy, and S. Marlow. A framework for event detection in field-sports video broadcasts based on SVM generated audio-visual feature model. case-study:soccer video. In *IWSSIP'04 - International Workshop on Systems, Signals and Image Processing*, 2004.

[SS98] Pasquale Savino and Fabrizio Sebastiani. Essential bibliography on multimedia information retrieval, categorization and filtering. In *Slides of the 2nd European Digital Libraries Conference Tutorial on Multimedia Information Retrieval*, 1998.

[SWY75] G. Salton, A. Wong, and C. Yang. A vector space model for automatic indexing. In *Communications of the ACM*, pages 613–620, 1975.

[THC03] Ming-Feng Tsai, Ming-Hung Hsu, and Hsin-Hsi Chen. Approach of information retrieval with reference corpus to novelty detection. In *Proceedings of the 12th Text Retrieval Conference (TREC 2003)*, 2003.

[Var99] Hal R. Varian. Economics and search. *SIGIR Forum*, 33(1):1–5, 1999.

[Voo04] E. M. Voorhees. Overview of TREC 2004. In *In Proceedings of the 13th Text Retrieval Conference (TREC 2004)*, pages 26–28, 2004.

[vR79] C.J. van Riisborgen. Information Retrieval (second edition). London: Butterworths, 1979.

[WK96] X. Wan and C.C. J Kuo. Color distribution analysis and quantization for image retrieval. In *Proceedings of SPIE: Storage and Retreival for Still Image and Video Databases IV, Vol 2670 pp. 8-16*, 1996.

[YS03] J. Ye and A. F. Smeaton. Aggregated feature retrieval for MPEG-7. In *European Conference in Information Retrieval 2003 LNCS Series 2633 Springer-Verlag*, 2003.

[yYH04] R. yan, J. Yang, and A. G. Hauptmann. Learning query-class dependent weights in automatic video retrieval. In *Proceedings of 12th ACM international conference on multimedia*, pages 660–667. ACM Press, 2004.

[ZCM02] Yi Zhang, Jamie Callan, and Thomas Minka. Novelty and redundancy detection in adaptive filtering. In *SIGIR '02: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 81–88, New York, NY, USA, 2002. ACM Press.