Aggregated Feature Retrieval for MPEG-7 via Clustering

Jiamin Ye
Centre for Digital Video Processing
Dublin City University
+353-1-7008563
jiaminye@computing.dcu.ie

Alan F. Smeaton
Centre for Digital Video Processing
Dublin City University
+353-1-7005262
asmeaton@computing.dcu.ie

ABSTRACT
In this paper, we describe an approach to combining text and visual features from MPEG-7 descriptions of video. A video retrieval process is aligned to a text retrieval process based on the TF*IDF vector space model via clustering of low-level visual features. Our assumption is that shots within the same cluster are not only similar visually but also semantically, to certain extent. Our TRECVID2002 and TRECVID2003 experiments show that adding extra meaning to a shot based on the shots from the same cluster is useful when each video in a collection contains a high proportion of similar shots, for example in documentaries.

Keywords
Video retrieval, MPEG-7, TRECVID, clustering

1. INTRODUCTION
A popular temporal representation of a video is to decompose the video into scenes and shots. Each shot can be annotated by a text description and further represented by a key frame, which can be described by low-level visual features. MPEG-7 is a generic standard used to encode information about multimedia content [1] and often, different MPEG-7 Descriptor Schemas are instantiated for different representations of a shot such as text annotations and visual features. Our work focuses on two main areas, the first is devising a method for combining text annotations and visual features into one single MPEG-7 description and the second is defining how best to carry out text and non-text queries for retrieval via a combined description. Our experiments in section 3 are concerned with the manual search task in TRECVID2002 and TRECVID2003. Section 4 summarises our main results.

2. AGGREGATED FEATURE RETRIEVAL
One of the challenges with video retrieval stems from the difficulty of combining different features that can be automatically extracted from video content and their detection accuracy. There are three types of features: (1) high-level text (e.g. ASR transcripts), (2) low-level audio/visual features (e.g. color, texture), (3) mid-level concepts (i.e. indoor, outdoor). Of the three types of features, text is the most precise representation of shot content and low-level features offer limited semantic information. Conceptual features were introduced to try bridging this semantic gap [2]. These are predefined concept labels, useful in limited domains such as a TV News anchorperson detector used to exclude shots containing anchorpersons before retrieval for shots takes place [3]. The number of such possible concepts however is so huge that to classify shots based on all concepts is impossible.

Our solution is to align a video retrieval process to a text retrieval process based on the TF*IDF vector space model via clustering of low-level visual features. Our assumption is that shots within the same cluster are not only similar visually but also semantically to a certain extent [5]. Our method maps the visual features of each shot onto a term weight vector via clustering. This vector is then combined with the original text features of the shot (i.e. ASR transcripts) to produce the final searchable index.

2.1 Index Preparation
An index for visual features is prepared by following three steps:

- Apply k-means shot-clustering to obtain clusters for each video. Features considered here include color histogram, dominant color and edge histogram as described in MPEG-7.
- Assign meanings to each cluster using a modified TF*IDF algorithm in which the indexed unit is replaced by a cluster.
- Use a simplified Bayesian approach to derive the text description of each shot based on its cluster meanings.

Having obtained a text description for shots via clustering, we aggregate this with the original term weight vector for each shot to create its final term weight vector using the Ordered Weighted Averaging operators and linguistic quantifiers [4].

2.2 Query Preparation
Two types of video query are under consideration in our work: (1) text-only, (2) non-text (i.e. consisting of an image/ video clip represented by a key frame). Low-level features are calculated for a non-text query and the final query is prepared as follows:

- Find the N most similar clusters to a given image example based on the low-level features.
- Take the term vectors of the N chosen clusters and aggregate them together to form a single query term vector.
- Aggregate the original text query and the derived text query (from the non-text query) to create the final query term vector.

SIGIR'04, July 25–29, 2004, Sheffield, UK.

2.3 Retrieval

Retrieval can be done in a straightforward way by calculating the dot products between the aggregated query term vector and the aggregated index vector for each shot and sorting the dot products in descending order.

3. EXPERIMENTS

Experiments were performed as the manual search task on both the TRECVID2002 and TRECVID2003 collections. A baseline system using ASR transcripts alone was built for each collection based on the TF*IDF model. The retrieval unit was defined as a shot and the ASR transcript that belongs to the shot was used solely in contributing to forming a term weight vector. Our evaluation objectives are to examine whether:

– an aggregated index built from visual features along with text features is useful in helping traditional text-only queries?

– an aggregated query derived from a non-text query along with the original text query useful in retrieval?

Retrieval results in Fig. 1 shows that an aggregated index yields marginally improved results over the baseline using the TRECVID2002 collection. The aggregated index does retrieve more relevant shots overall but could not improve their rankings. The introduction of cluster meanings to its member shots not only enhances the shot meanings but also weakens them to some extent. It is also shown that an aggregated query gave slight better performance over the baseline and it is topic-specific. If no correct clusters can be found, the derived query term vector will be poorly formed.

Fig. 2 shows that there is no significant improvement in using an aggregated index for the TRECVID2003 collection. The TRECVID2003 collection mostly contains videos of CNN and ABC broadcast TV news. Except for anchorperson, commercial and sports shots, there are few shots remaining that are similar visually and semantically within each news programme. Clustering shots within a single news programme is thus ineffective compared to clustering within a programme in the TRECVID2002 collection. The TRECVID2002 collection has videos of documentaries from the 1940s to 1960s and there is a good amount of shots within a video which are not necessarily the same but are similar both visually and semantically.

4. CONCLUSIONS

A typical video retrieval approach is to search different MPEG-7 descriptions separately and to combine ranked results from the different searches using a sum weighted method. Our approach attempts to integrate the different descriptions into the index and query preparation stages and no combination of ranked results is required. This is useful when each video in a collection contains a high proportion of shots that are similar visually and semantically to some extent. Adding meaning to a shot based on the shots that are around it is an effective method for video retrieval when each video in the collection has low proportion of similar shots (i.e. TV news programmes).

5. ACKNOWLEDGMENTS

The support of the Informatics Directorate of Enterprise Ireland is gratefully acknowledged. This work is based upon work supported by the IST programme of the EU in the project IST-2000-32795 SCHEMA.

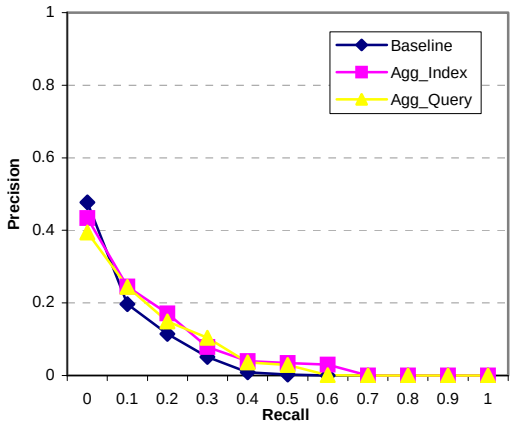

Figure 1. Precision at recalls for TRECVID2002

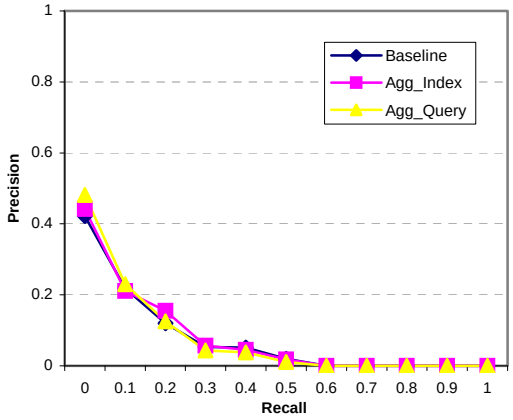

Figure 2. Precision at recalls for TRECVID2003

6. REFERENCES

[1] Martínez, J.M. (ed.) MPEG-7 Overview (v.9.0), MPEG/N5525, Pattaya, March 2003. Last visit: 12 Feb 2004, http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm.

[2] Smeaton, A. and Over, P. The TREC 2002 Video Track Report. In *The 11th Text Retrieval Conference*, Gaithersburg, 5 November 2002, 69-85.

[3] Westerveld, T., Vries, A.P., Ianeva, T., Boldareva, L. and Hiemstra, D. Combining Information Sources for Video Retrieval. In *the 12th Text Retrieval Conference*, Gaithersburg, November 2003. Last visit: 12 Feb 2004, http://www-nlpir.nist.gov/projects/tvpubs/papers/lowlandsteam.paper.pdf

[4] Yager, R. A Hierarchical Document Retrieval Language. *Information Retrieval*, Vol.3, No.4, December 2000, 357-377.

[5] Ye, J. and Smeaton, A. Aggregated Feature Retrieval for MPEG-7. Poster In *the Proceedings of 25th European Conference on IR Research (ECIR)*, Pisa, Italy, April 2003, 563-570.