Object-Based Access to TV Rushes Video

Alan F. Smeaton, Gareth J.F. Jones, Hyowon Lee, Noel E. O'Connor, and Sorin Sav

Centre for Digital Video Processing & Adaptive Information Cluster,
Dublin City University, Glasnevin, Dublin 9, Ireland.
Alan.Smeaton@dcu.ie

1 Introduction

Recent years have seen the development of different modalities for video retrieval. The most common of these are (1) to use text from speech recognition or closed captions, (2) to match keyframes using image retrieval techniques like colour and texture [6] and (3) to use semantic features like "indoor", "outdoor" or "persons". Of these, text-based retrieval is the most mature and useful, while image-based retrieval using low-level image features usually depends on matching keyframes rather than whole-shots. Automatic detection of video concepts is receiving much attention and as progress is made in this area we will see consequent impact on the quality of video retrieval. In practice it is the combination of these techniques which realises the most useful, and effective, video retrieval as shown by us repeatedly in TRECVid [5].

For many types of query we seek video which contains an *object* of interest such as a car, a building or an animal, something where the background is of no importance. Here we introduce a technique we have developed for object-based video retrieval. We outline the processes involved in analysing and indexing video to support this and we present an interactive system to support user searching using objects and/or using matching of whole keyframes.

The data used in our work is 50 hours (c. 4.5M frames) of rushes video provided by BBC as part of the TRECVid evaluation in 2005. Rushes is a term used to refer to raw video footage which is unedited and contains lots of redundancy, overlap and "wasteful" material. Shots tend to be much longer than in post-produced video and it generally contains a lot of re-takes, bloopers and content where nothing much happens. It is very similar to home movies or personal video material since it often contains camera shake and re-focus, and little or no dialogue. The task for participants in this track in TRECVid 2005 was to explore how to develop techniques to automatically analyse such video given that there is no text dialogue to work with and to build systems which allow users who know nothing about the content of the data to navigate through it with some information need in mind.

2 Object-Based Video Retrieval

In work reported elsewhere [2] we developed a video retrieval and browsing system which allowed users to search using the text of closed captions, using the keyframe for locating similar keyframes in terms of colour, texture and edges, and using the occurrence (or non-occurrence) of a set of pre-defined video objects. The content used was several seasons of the Simpsons TV series and the video objects corresponded to the faces of the 10 major characters in the series, Homer, Bart, Marge, etc. We evaluated the ways in which different video retrieval modalities (text search, image search, object search) were used [3] and we concluded that certain queries can benefit from using objects as part of their search, but this is not true for all query types.

In moving from object detection and retrieval on synthetic video (e.g. the Simpsons) to object retrieval on natural video as in [4], we are faced with a problem of object segmentation. This is hard because in video objects can deform and turn, cameras can move, objects can become occluded when other objects move in front of them, lighting conditions can change, and so on. Nevertheless we have developed a semi-supervised technique for object segmentation based on an RSST region segmentation which requires the user to indicate some regions both within and outside the contours of the object to be segmented and this can be done easily with two mouse-strokes, one inside and one outside the object [1]. This approach is feasible for the BBC rushes video corpus in comparison with the amount of manual effort currently placed on video annotation, and we've segmented objects from this corpus of video data in order to support object retrieval. Our retrieval system also supports whole keyframe based retrieval where a number of images can be used as the query and in object based retrieval a number of example objects can be used as the query. These two approaches can also be combined into one query for retrieval and our interest is in seeing under what circumstances users find object-based retrieval to be useful.

We used a standard approach to shot boundary determination, comparing adjacent frames over a certain window and using low-level colour features, in order to determine boundaries [5]. We detected 8,717 shots (a rate of 174 keyframes per hour) for the 50 hours and for each of these we automatically extracted a single keyframe by examining the whole shot for levels of visual activity using features extracted directly from the encoded video. The rationale for this is that the approach of choosing the first, last or middle frame as the keyframe would be quite inappropriate given the amount of “dead” time there is in shots within rushes video. Much of the unusable video footage in rushes is there because the camera is left running while the main action of the shot is prepared and then takes place. In rushes footage the camera is left running in order to ensure the action, whatever that may be, is not missed. Thus our approach to automatic keyframe selection based on choosing the frame where the greatest amount of action is happening, seems to make sense, although this is certainly a topic for further investigation.

Each keyframe was manually examined to determine if there was a single dominant object present and if so it was segmented from its background using our semi-automatic tool [1] described in the previous section [1] which yielded 1,210 objects. Once segmentation was completed we extracted features for keyframes using global MPEG-7 colour and texture features for whole keyframes and dominant colour, homogeneous texture and shape compactness MPEG-7 features for objects. We then pre-computed two $8,717 \times 8,717$ matrices of keyframe similarities using the two image features for the whole keyframe, and three $1,210 \times 1,210$ matrices of similarities between objects in those keyframes using object features.

For retrieval we cannot assume that the user knows the archive's content since rushes content is not catalogued in any way. In order to begin a user's retrieval we ask the user to locate one or more external images using some other image searching resource. The aim here is to find one or more images, or even better one or more video objects, which can be used for searching. In our experiments our users use Google Image Search to locate such external images but any image searching facility could be used. Once external images are found they are indexed in the same way as keyframes in terms of colour and texture for the whole image and the user is also allowed to segment one object in the external image if s/he wishes. This is done in real time.

At search time the user indicates which visual characteristics in each query image are important — colour or texture in the case of the whole image, or colour, shape or texture in the case of an object. The set of query images is then used for retrieval and the user is presented with a list of keyframes from the archive. The similarity between these and the user query is a combination of image-image similarity (using colour and texture) and object-object similarity (using colour, shape and texture). For the 1,210 of 8,717 keyframes where there is a segmented object present the object is highlighted when the keyframe is presented. The user browses these keyframes and can either play the video, save the shot, or add the keyframe (and its object, if present) to the query panel and the query-browse iteration continues. The overall architecture and a sample screen taken from the middle of a search is shown as Figure 1 where there are 3 query images (the first is a whole keyframe added from within the collection as relevance feedback; the second and third are external images and the user has segmented objects in these), 5 pages of search results and 4 saved keyframes. Objects appearing in frames with segmented objects, either in the query panel or search result, are outlined in red and the facets of the query images which are to be used for the search (colour, texture, object shape) are shown in the query panel.

3 Experiments and Plans

We now have two versions of our system allowing us to explore how useful video objects are in video browsing and search. One supports image similarity based only on whole keyframes, while the other supports object-object similarity as well as whole keyframe matching. A small user experiment with 12 users is being run to compare the performance of the two systems for a number of searches. We are using a Latin squares design for rotating the order of users for topics, and systems. Since there was no formal evaluation task for BBC rushes at TRECVid 2005 (it was an exploratory task only), we take a "do-it-yourself" approach to formulating search topics and performing relevance assessments. We will share

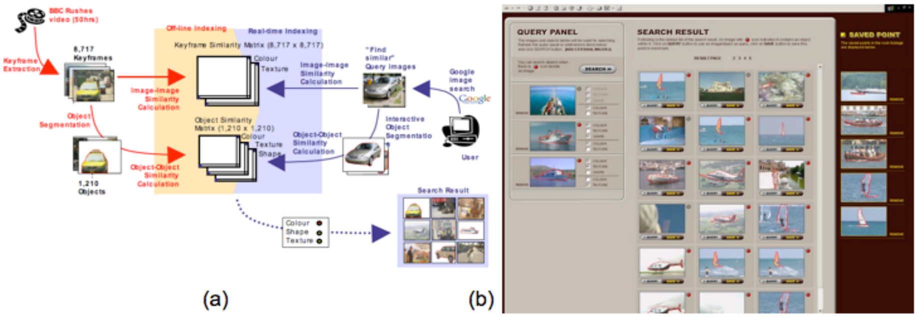
BBC Rushes video (50hrs)
Keyframe Extraction
8,717 Keyframes
Object Segmentation
1,210 Objects
Image-Image Similarity Calculation
Object-Object Similarity Calculation
Off-line Indexing
Real-time Indexing
Keyframe Similarity Matrix (8,717 x 8,717)
Colour
Texture
Object Similarity Matrix (1,210 x 1,210)
Colour
Texture
Shape
Image-Image Similarity Calculation
Object-Object Similarity Calculation
"Find similar" Query images
Google image search
Google
Interactive Object Segmentatio
User
Colour
Shape
Texture
Search Result
(a)
(b)
QUERY PANEL
SEARCH
SEARCH RESULT
SAVED POINT

Fig. 1. (a) System architecture and (b) sample screen for video rushes search system

our topics, pool retrieved shots and share relevance judgments with at least one other TRECVid participating group in a a mini-TRECVid for the BBC rushes data. Search topics are taken from a log of actual search topics from a broadcaster's archive to which we have access.

The use of video objects in searching offers interesting potential to expanding the set of possible modalities for video search and browsing but our dependency on using objects from single keyframes is limiting. We index objects in a keyframe rather than in a shot, and as we know during a shot an object can turn or deform, and the camera and/or the object can move, all yielding different object representations. To overcome this we would like to track an object throughout a shot and to index each *instance* of the object throughout the frame. Although this is very ambitious it would reduce dependency on keyframes rather than whole shots and is planned as further work.

References

1. Adamek, T. and O'Connor, N. (2003). Efficient Contour-based Shape Representation and Matching. MIR 2003 - 5th International ACM SIGMM Workshop on Multimedia Information Retrieval, Berkeley, CA, 7 November 2003.
2. Browne, P. and Smeaton, A.F. (2004). Video Information Retrieval Using Objects and Ostensive Relevance Feedback. In Proceedings of SAC 2004 - ACM Symposium on Applied Computing, Nicosia, Cyprus, 14-17 March 2004.
3. Browne, P. and Smeaton A.F. (2006). A Usage Study of Retrieval Modalities for Video Shot Retrieval. *Information Processing and Management*, (in press),
4. Sivic, J. and Zisserman, Z. (2003) Video Google: A Text Retrieval Approach to Object Matching in Videos. Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003).
5. Smeaton, A.F., Kraaij, W., and Over, P. (2004). 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, Avignon, France, 26-28 April 2004.
6. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A. and Jain, R. (2000). Content-Based Image Retrieval at the End of the Early Years. IEEE Trans. Pattern Anal. Mach. Intell. **22(12)**, pp. 1349-1380.