Efficient and Effective Similarity-based Video Retrieval*

Ilaria Bartolini
DEIS, University of Bologna, Italy
i.bartolini@unibo.it

Corrado Romani
DEIS, University of Bologna, Italy
corrado.romani@unibo.it

ABSTRACT

The retrieval of videos of interest from large video collections is a main open problem which calls for the definition of new video content characterization techniques in term of both visual descriptors and semantic annotations. In this paper, we present an efficient and effective video retrieval system which profitably exploits the functionalities offered by a semantic-based automatic video annotator using video shots similarity to suggest relevant labels for the videos to be annotated. Similarity queries based on semantic labels and/or visual features are implemented and experimentally compared on real data in order to measure the retrieval contribution of each type of video content information.

1. INTRODUCTION

Efficient and effective video retrieval is one of the most challenging tasks of Information Retrieval [6]. Some of the systems that have been proposed in the last few years rely only on visual features for indexing data (e.g., [9]), thus suffering the drawback coming from the semantic gap existing between the user subjectivity notion of similarity and the one implemented by the system. Video indexing based on semantic annotations [7] seems to be a better option for the user than visual features, although they are still used as post-filtering of the semantic result. Finally, some systems focus on a multi-modal retrieval paradigm [6] allowing both semantic concept- and feature-based retrieval.

In this paper we present an effective and efficient video retrieval system which conjunctively exploits the accuracy of automatically provided semantic-based hierarchical video annotations and the similarity between video scenes in terms of visual features in order to satisfy user expectations. The annotation process uses labels of pre-annotated key frames, following the key idea to suggest, for a given key frame, those tags which are assigned to key frames that are *similar* to it. In this way we are able to attach tags to video shots that we then propagate at the whole video level, so as to obtain semantic indices for a hierarchical (two-level) browsing platform. In addition to the above described query paradigms, we also support a combined query modality useful to enlarge the query results when semantic annotations are incomplete. Preliminary results on the TRECVID benchmark [10] provide evidence of the accuracy of the video annotation and the precision of the video retrieval.

*This work is partially supported by the CoOPERARE MIUR Project.

SISAP 2010, September 18-19, 2010, Istanbul, Turkey.

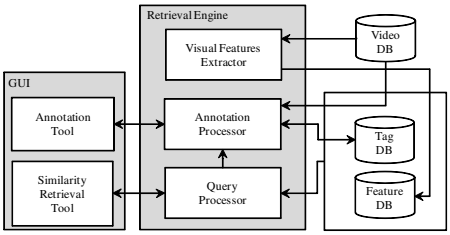


Figure 1: Retrieval architecture.

2. ARCHITECTURE AND PRINCIPLES

Figure 1 provides an overview of our video retrieval system. For each video in the *Video DB*, the *Visual Features Extractor* analyzes its shots and extracts a set of visual features from each key frame [3]. Each key frame is automatically segmented into a set of homogeneous regions which convey information about color and texture [2]. Extracted features are saved into the *Feature DB* and then indexed with an implementation of the M-tree metric index [4] to provide an efficient access and speed up the retrieval phase.

For each shot, the *Annotation Processor* assigns labels depending on its visual features. After processing all the shots, the annotator selects the most appropriate tags for the whole video and saves all the information into the *Tag DB*. An inverted file is maintained in order to guarantee an efficient tag-based retrieval. The tagging phase exploits the Imagination system [1] and uses a set of pre-annotated images as a knowledge base.

In details, given a key frame to be labelled, the annotator uses its visual features and, by exploiting the M-tree index, suggests those tags that are assigned to key frames in the knowledge base that are similar to the target one. In this way, a set of semantic concepts is attached to each representative key frame of a scene. Only terms recurring in the majority of key frames are selected as suitable concepts to describe the whole shot sequence. To avoid producing an overwhelming number of tags for each shot, only the most frequent tags retrieved for each key frame in the sequence are maintained. Shot tags are useful to browse sequences across different videos, but they could be too specific to in-

dex a whole video, especially if this contains a wide range of different visual content. A simple criterion to select video tags from the set of shot tags is to weigh every tag depending on its frequency and on the length of the shot it is associated to: concepts extracted from long shot sequences and/or that appear in several shots are probably more relevant, to describe the content of a whole video, than concepts occurring rarely or in short sequences.

At query time, the user submits her requests to the system; the *Query Processor* manages such requests in order to return the videos (respectively shots) of interest. Three main query paradigms are supported: *keyword*-based (*KS*), *feature*-based (*FS*) and *keyword&feaure*-based (*KFS*) searches, respectively. The *KS* paradigm is the easiest and most popular query modality used by traditional search engines, where the user enters a set of keywords as query semantic concepts. Videos/shots are selected by the query processor by applying a *co-occurrence* search on the tag DB. The search provides the set of videos (resp., shots) that share at least one tag with the input set. We rank the returned objects on the base of the co-occurrence value and return the top-k ones.

With the *FS* modality, the user is looking for those shots whose representative key frames are similar to an input query image. A *nearest-neighbors* (NN) search is performed on the Feature DB. Since a shot can be represented by more than a single key frame, we compute a k'-NN search, with ($k' \simeq 2k$) in order to derive a sufficient number of distinguishing shots. Candidate shots are ranked on the EMD distance we used to compare key frames [2] and the top-k are returned.

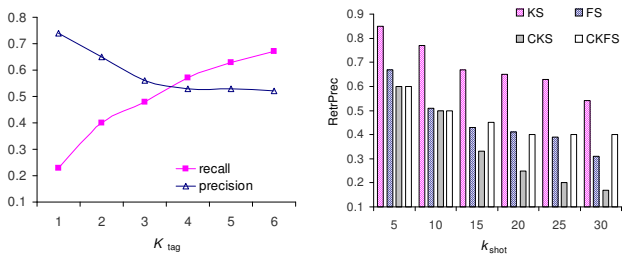
Finally, *KFS* queries combines *KS* and *FS* by returning shots in the intersection of both *KS* and *FS* results first, followed by shots in the *KS* list only and, finally, by shots in the *FS* result only. This query modality is particularly convenient when keyword-based search involves concepts which are poorly represented in knowledge base, i.e., shots annotated with that concept are less than k. In this case, by only applying *KS* we would not be able to return k shots. If *KS* is able alone to provide the desired number of objects, adding features in the query process is quite pointless since it just re-ranks the same result set.

3. EXPERIMENTAL VALIDATION

We implemented our retrieval system in Java JDK 6.0 and tested it on the TRECVID benchmarks [10]. We used the knowledge base of the *TRECVID-2007 High-Level Feature* task [10] which consists of 110 videos (50 hours total length), 21500 key frames, and 9000 shot cuts. Each shot is described by means of tags coming from 36 semantic concepts.

We first evaluate the accuracy of the annotator in term of classical *precision* (number of relevant retrieved tags over the returned labels) and *recall* (number of relevant retrieved tags over the total number of relevant tags for the shot) metrics over a set of provided testing videos which come with a ground truth for evaluation purposes. Then we tested the video retrieval effectiveness over a set of random selected queries (as described in the following) by means of the *retrieval precision* metric, i.e., the number of relevant shots returned over the number of returned shots.

Figure 2 (a) shows the average annotation precision and recall values, when varying the number of predicted tags. As one can observe, the annotator performs quite well, representing a good starting point upon which an effective video retrieval system can be built.



(a) (b)

Figure 2: Average accuracy of the annotator in terms of (a) precision and recall vs. number of predicted tags and (b) retrieval precision of the query processor vs. number of returned shots.

In Figure 2 (b) the average accuracy of the video retrieval vs. the number of returned shots is shown. The query workload consisted of 7 different queries: *outdoor scenes*, *scenes containing persons*, *faces*, *roads*, *sky views*, *vegetation*, and *waterscape/waterfront views*. We ran keyword-based (KS) similarity searches and evaluated the annotations of the top-k shots. Then we ran feature-based similarity search (FS) by computing the top-k shots which are most similar to the query image corresponding to each semantic query concept. KS performs very well (e.g., about 85% of precision for the first 5 shots and 65% when 25 shots are retrieved); the retrieval precision for FS is satisfactory, although quite worse than KS (as expected because of the semantic gap problem).

Finally, we tested the mixed query paradigm, by assuming a *poor* annotation scenario: the concept "Car" in our ground truth appears in 31 shots, but our annotator predicted it for 7 shots. We denote with CKS the "Car keyword-based search", while with CKFS the "Car keyword/feature-based search". It is immediate to derive from Figure 2 (b) the contribution of the visual features to CKS: CKFS is indeed able to enlarge the cardinality of the relevant results by providing 8 more correct matches, while maintaining acceptable level of precision values, especially when $k_{shot} > 10$.

4. REFERENCES

[1] I. Bartolini and P. Ciaccia. Imagination: Accurate Image Annotation Using Link-analysis Techniques. In *AMR 2007*, pages 32–44, 2007.

[2] I. Bartolini, P. Ciaccia, and M. Patella. Query Processing Issues in Region-Based Image Databases. In *Knowledge and Information Systems (KAIS)*, 2010. To appear.

[3] I. Bartolini, M. Patella, and C. Romani. SHIATSU: Semantic-Based Hierarchical Automatic Tagging of Videos by Segmentation using Cuts. In *AIEMPro 2010*. To appear.

[4] P. Ciaccia, M. Patella, and P. Zezula. M-tree: an Efficient Access Method for Similarity Search in Metric Spaces. In *VLDB 1997*, pages 426–435, 1997.

[5] C. Diou, G. Stephanopoulos, N. Dimitrou et al. VITALAS at TRECVID-2009. In *TRECVID 2009*, pages 16–17, 2009.

[6] P. Geetha, and V. Narayanan. A Survey of Content-based Video Retrieval. In *Journal of Computer Science*, 4(6), pages 474–486, 2008.

[7] A. G. Hauptmann, M. G. Christel, and R. Yan. Video Retrieval Based on Semantic Concepts. In *Proceedings of the IEEE*, 4(96), pages 602–622, 2008.

[8] C.-W. Ngo, Y.-G. Jiang, X.-Y. Wei et al. VIREO/DVMM at TRECVIDE 2009: High-Level Feature Extraction, Automatic Video Search and Content-Based Copy Detection. In *TRECVID 2009*, pages 16–17, 2009.

[9] T. N. Shanmugam, and P. Rajendran. An Enhanced Content-Based Video Retrieval System Based on Query-Clip. In *International Journal of Research and Reviews in Applied Sciences*, 1(3), pages 236–253, 2009.

[10] TRECVID Video Retrieval Evaluation: http://trecvid.nist.gov.