Region Feature Based Similarity Searching of Semantic Video Objects

Di Zhong and Shih-Fu Chang
Image and Advanced TV Lab, Department of Electrical Engineering
Columbia University, New York, NY 10027, USA
{dzhong, sfchang}@ee.columbia.edu

Abstract

New video representations based on semantic objects (e.g., MPEG-4) provide great potential for content-based video searching. In this paper, we present an efficient query model for similarity searching of video objects based on localized region features and spatial-temporal structures. The query model is based on an existing framework on region-level video query, but addresses several new issues like tight spatio-temporal relationships and multi-level feature hierarchy. We also include experiment results to demonstrate the effective performance of the proposed approach.

1. Introduction

Content-based image and video retrieval has been studied by many researches in the recent years. Most of existing systems provide methods for content-based retrieval of images and videos by using query examples and sketches [1,5].

Our previous work, VideoQ[1], is a content-based video searching system which allows users to search video clips based on a rich set of region visual features and spatio-temporal relationships. It uses local region features directly for video search. While region segmentation and region-level feature extraction can be fully automatic, further efforts have to be made to support high level semantic queries. In recent MPEG-4 standard, an object-based coding representation for audio-visual data is proposed. MPEG-4 objects are *semantic* objects corresponding to meaningful real-world objects. This introduces one more description level on top of the region-level indexes.

While the matching of multiple regions with spatial-temporal relationships in VideoQ can be applied to similarity search of semantic video objects, there are several new critical issues. First, underlying regions of a video object are tightly connected with each other, and thus similarity matching of spatial-temporal structures become more important and should be performed more precisely. Secondly, the problem of partial matching between query regions and object regions stored in the feature library need to solved. In the previous work, query regions can be a subset of regions of a matched object (or video clip). However, a matched object has to contain matches for all query regions. This implicitly limits the number of regions that a user can use to define an object. Here we propose a efficient partial matching method so that users can describe query objects in detail with more regions. Finally, there are efficiency concerns, especially when using a rich set of temporal and spatial structure features. It is time-consuming to compute and match them in query time. They need to be pre-computed and properly indexed for efficient query processing.

In this paper, we extend our prior work, VideoQ[1], to develop a new content-based search system for semantic video objects. The unique features of the system include:

- a rich set of visual features at both the object and region levels, including color, shape, texture, motion trajectory and temporal information.
- effective integration of global and localized feature matching, as well as the measurement of various spatial-temporal structures (directional, topological, temporal).
- a flexible and convenient query interface which allows users to easily create a query object by using example objects or by drawing on the canvas.

The paper is organized as follows. In section 2, we briefly discuss a region segmentation and tracking system which is used to extract salient feature regions of video objects. In Section 3, we present an efficient query model for similarity search of video objects based on feature region and various spatial-temporal structures among them. Experiments and analysis are given in Section 4. Finally, section 5 provides the conclusion.

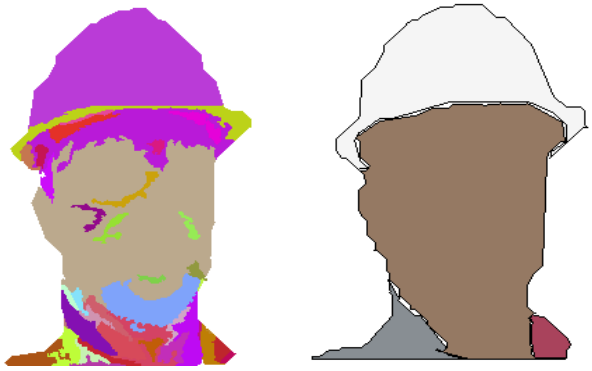
2. Extraction of Salient Feature Regions

We modified our algorithms developed in the AMOS[1]

1. the binary software of AMOS is available to the public at http://www.ee.columbia.edu/advent/

system[6] to extract salient regions from video objects. The AMOS system is an active object segmentation and tracking system for general video sources. It allows users to identify a semantic object in the starting frame of a video shot and then track the object from frame to frame through a region decomposition, projection and aggregation process.

Although in the AMOS system, the underlying regions of video objects are already created and tracked during the object segmentation process, these regions usually have small sizes and short durations in order to achieve accurate object boundaries. Thus extra efforts (e.g., grouping or user interaction) are required to create salient feature regions that can facilitate efficient object retrieval. As shown in **Figure 1,** regions extracted for the searching purpose are usually larger and fewer in contrast to those extracted for the coding purpose.



a. small regions for accurate object segmentation

b. salient regions for feature based object search

Figure 1. Difference between underlying regions for object segmentation and object search

To allow users to define higher level "semantic" regions which may not be uniform in terms of visual features, the new system also provides convenient interactive functions to help users to manipulate or create desired feature regions. Automatic tracked regions can be further merged semi-automatically. Users can click on two neighboring regions at any frame to merge them in all the frames where these two regions exist. In general, this allows flexible user control to create desirable region segmentation and salient features within a video object.

After the above salient region segmentation process, each segmented region is stored as a list of masks (approximated using polygons) with its starting and ending frame number in a video sequence. Video objects are stored in the same way. All these data will be used for the following feature extraction process.

3. Region-Based Object Query Model

In this section, we present an efficient query model for similarity search of video objects based on localized region features and various spatial-temporal structures among them. It requires an integrated matching of visual features and spatial-temporal structures. We first discuss the feature extraction process.

3.1. Visual Features

Similar to those in the VideoQ system, a large set of visual features of video objects and their underlying regions are computed and stored in a visual feature library. Major features include motion trajectory, mean color, TAMURA texture and shape descriptors. Detailed descriptions of the features can be found in [1]. Note that the above feature data can be computed at all sampled frames or only frames where differences exceed certain thresholds. The later approach extracts features at significant events only and allows flexible temporal variations.

Each video object has a unique *ObjectID*. Similarly each region also has a unique *RegionID*. A feature vector is stored together with its corresponding *ObjectID* and *RegionID*. These *ID's* are used as index in the following object matching and retrieval process.

3.2. Spatial-Temporal Structure Features

Given a set of regions of an object, the structure features can be derived from their spatial and temporal positions and boundaries. These features need to be pre-computed and properly stored so that they can be quickly accessed and examined in the similarity matching process. In the following, we proposed two spatial and one temporal structure features (**Figure 2**).

There are several different ways to compare spatial structures, such as 2D-Strings [8] and spatial-temporal logics [9]. We use a relatively fast and simple method, the spatial-orientation graph, to represents spatial relationships as edges in a weighted graph [2]. This graph can be easily constructed by computing the orientation of each edge between the centroids of a pair of query regions, and stored as a n*(n-1)/2-dimension feature vector, where *n* is the number of query regions (**Figure 2a**). As the spatial-orientation graph cannot handle the "contain" relationship, we extend it with a topological graph (**Figure 2b**), which defines the contain relation between each pair of regions with three possible values: contains(1), is-contained(-1) or not containing each other(0). Similarly, the temporal graph (**Figure 2c**) defines whether one region starts before(1), after(-1) or at the same time(0) with another region. Note here for simplicity we only define the temporal order

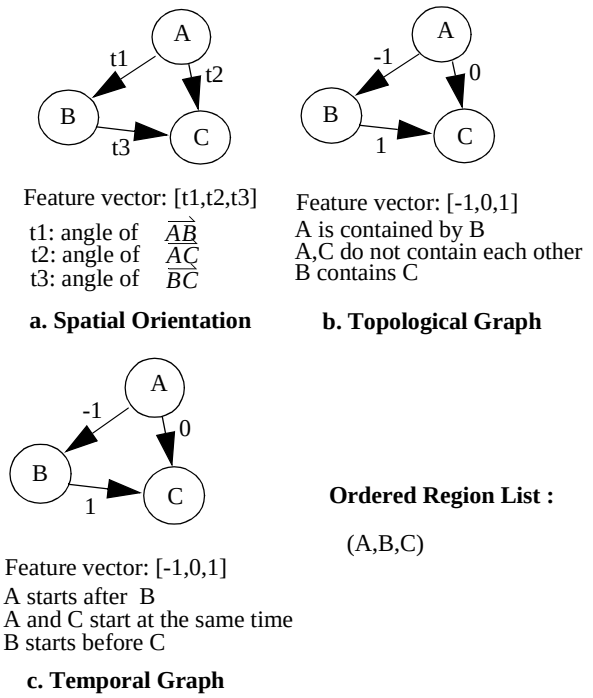


Figure 2. Examples of the three spatio-temporal relationship graphs

according to the first appearing time (or starting time) of each region. By taking the ending time, a more complicated temporal relation graph can also be generated.

Given the large amount of feature data, indexing is certainly important for efficient search. Separated or integrated indexes can be built on each feature vector. Efficient indexing and retrieval models in this type of multi-region, multi-feature similarity searching environment has been studied in [7]. In the remaining parts of this paper, we focus on visual features and their similarity matching.

3.3. Query Model

Given a query object (composition of a set of query regions), there are generally two searching approaches. One is to directly match query object against objects in the database. R-Tree based search methods is proposed in [4]. However, R-tree is not suitable for indexing of a large set of high-dimension features.

The other searching approach is to first find a matching region list for each query region based on visual features, and then "join" these region lists to find the best matched video objects by combining visual and structure similarity metrics. We used this approach in our earlier work, VisualSEEK [5]. In [3], Li and Smith further proposed a query-

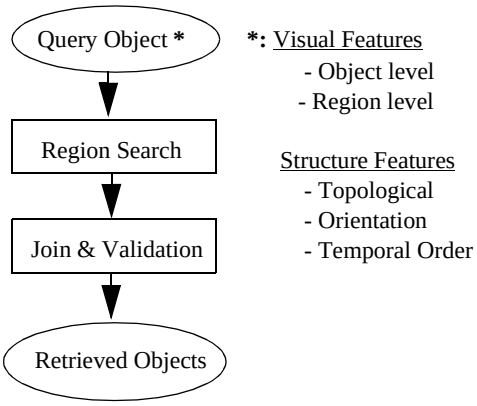


Figure 3. The video object query model

ing scheme which divides a composite query into a sequential of representation of sub-queries. A fast dynamic programming method is then used to retrieve the best matches for a composite object. However, sequentialization of a large set of visual and structure features is difficult. In this paper, we use a parallel query and join scheme which supports partial matches. The object searching diagram is shown in **Figure 3**.

Given a query object with N regions, the object searching process consists of two stages. The first stage, *region search*, is to find a candidate region list from the database for each query region. The second stage, *join & validation*, is to join the candidate region lists to produce the object candidate list and to compute the final global distance measure. The detailed procedure is given as follows:

1) For every query region, find a candidate region list based on the weighted sum (according to the weights given by users) of distance measures of different visual features (e.g., shape or trajectory). All individual feature distances are normalized to [0,1]. Only regions with distances smaller than a *threshold* are added to a candidate list. Here the *threshold* is a pre-set value used to empirically control the number or percentage of objects the query system will return. For example, a threshold 0.3 indicates that users want to retrieve around 30 percent of video objects in the database[1]. The threshold can be set to a large value to ensure completeness, or a small value to improve speed.
2) Sort regions in each candidate region list by their *ObjectID's*.
3) Perform join (outer join) of the region lists on *ObjectID* to create a candidate object list. Each candidate object

1. Assume the feature vectors of the video objects in the database have normal distribution in feature spaces.

in turn contains a list of regions. A "NULL" region is used when :

- a region list doesn't contains regions with the *ObjectID* of a being-joined object
- a region appears (i.e. matched) more than once in a being-joined object

4) Compute the distance between the query object and each object in the candidate object list as follows:

$$D=w_0\sum_i FD(q_i, r_i) + w_1 SD(sog_q, sog_o) + w_2 SD(topo_q, topo_o) + w_3 SD(temp_q, temp_o)$$

where q_i is the *i*th query region. r_i is the *i*th region in a candidate object. FD(.) is the feature distance between a region and its corresponding query region. If r_i is NULL, maximum distance (i.e., 1) is assigned. sog_q (spatial orientation), $topo_q$ (topological relationship) and $temp_q$ (temporal relationship) are structure features of the query object. sog_o, $topo_o$ and $temp_o$ are retrieved from database based on *ObjectID*, *RegionID* and temporal positions. When there is a NULL region (due to the above join process), the corresponding dimension of the retrieved feature vector will have a NULL value. SD(.) is the L1-distance and a penalty of maximum difference is assigned to any dimension with a NULL value.

5) Sort the candidate object list according to the above distance measure D and return the result.

4. Experiment Results

A prototype system has been developed to demonstrate and evaluate our proposed visual similarity searching method which integrates matching of localized feature matching and spatial-temporal structures. We created a semantic video object database with 104 video objects. This database is used as the testbed of our object searching system. Many different types of objects are included, such as people, sports, animal, flowers and transportation. About 500 salient regions and their visual features are extracted and stored in the database.

The query interface is shown in **Figure 4**. The system allows users to compose a visual query by drawing regions from scratch or loading underlying regions of an example object from the database. Once a set of regions have been created on the canvas, users can assign and change their visual attributes to form a query object. These attributes include size, shape, spatial positions, color, texture, motion trajectories, temporal position and temporal shape changes (e.g., rotation).

With all these properties, users can create a query object with specific visual features and spatial-temporal structures, and then choose a set of weights to start the search process. These weights define how important different region features and object features are in the similarity matching functions. Users also have the option of including matches of spatial and/or temporal structures in the query.

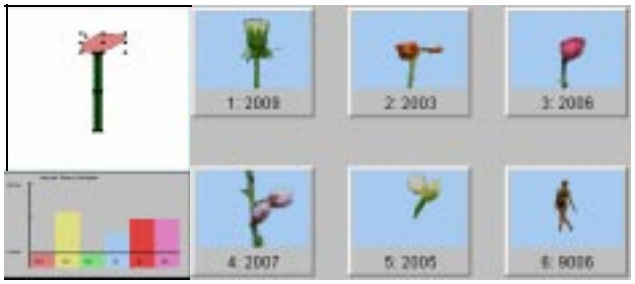
Figure 4. The user interface of the query system

Our initial query experiments have shown encouraging results. The matching of structure features has been proved to be critical in finding objects with multiple parts and special relationships (e.g. human body). Four query examples are shown in **Figure 5.** In the first example (**Figure 5a**), we are trying to find flowers by specifying two regions with color, shape and spatial relationship. Higher weight is put on the shape feature. Our experiments show that the shape matching of the stem part plays a key role in successfully finding flowers. In the second example (**Figure 5b**), we define a more complex query object, which has four parts with specific spatial relationship. While color and shape features are important in finding similar feature regions for each part, the matching of spatial structure eliminates false candidates and brings correct objects to the top of the return list. **Figure 5c** shows an example of partial matching. Users draw a face with two eyes. In the feature library, small regions like eyes are usually not extracted. Using the outer join and partial match method, we are able to find similar faces including those do not have extracted eye regions (results except the first one shown in **Figure 5c**). **Figure 5d** shows an example of trajectory matching. The retrieved objects are moving to the left-bottom direction on the screen.

5. Conclusion

A region-based object query model which effectively combines local region-level features and spatial-temporal structures is presented in this paper. A prototype content-

based video object searching system is developed. The system provides a powerful interface for users to create query objects with various visual features and various spatial-temporal structures. Initial experiments have shown promising results and great promise in developing advanced video search tools for semantic video representations such as MPEG-4.



(a) flowers

(b) body

(c) face

(d) trajectory

Figure 5. Video object query examples

In future work, building upon the visual searching tools of video objects, we will explore semantic object level video search. We are also studying application of the object search tools to MPEG-4 scene descriptions (i.e., BIFS).

Acknowledgment

We thank the Hot Shots & Cool Cuts Inc. and the Action, Sports & Advantage Inc. for providing some video content for the research.

References

[1] S.-F. Chang, W. Chen, H. Meng, H. Sundaram, and D. Zhong, "*VideoQ*: An Automated Content-Based Video Search System Using Visual Cues", ACM 5th Multimedia Conference, Seattle, WA, Nov. 1997.

[2] V.N. Gudivada and V.V. Raghavan, "Design and Evaluation of ALgorithms for Image Retrieval by Spatial Similarity", ACM Transaction on Information Systems, Vol.13,No,2,April 1995, pp.115-144.

[3] C.-S. Li, J.R. Smith, L. Bergman and V. Castelli, "Sequential Processing for Content-based Retrieval of Composite Objects", SPIE Storage&Retrieval of Image/Video DB, 1998.

[4] E.G.M. Petrakis and C. Faloutsos, "Similarity Searching in Large Image Database", TR# 3388, Dept. of Computer Science, University of Marryland.

[5] J.R.Smith and S.F.Chang, "VisualSEEK: a Fully Automated Content-based Image Query System", ACM Multimedia 96, Boston, MA, Nov 1996.

[6] D. Zhong and S.-F. Chang, "AMOS: an Active System for MPEG-4 Video Object Segmentation", 1998 International Conference on Image Processing, October 4-7, 1998, Chicago, Illinois, USA.

[7] H. Sundaram, S.-F. Chang "Efficient Video Sequence Retrieval in Large Repositories," Proc. SPIE Storage and Retrieval for Image and Video Databases VII, San Jose CA, Jan 1999.

[8] S.-K. Chang, Q. Y. Shi, and C. Y. Yan, "Iconic indexing by 2-D strings", IEEE Trans. Pattern Anal. Machine Intell., 9(3):413-428, May 1987.

[9] A. D. Bimbo, E. Vicario and D. Zingoni, "Symbolic description and visual querying of image sequences using spatio-temporal logic", IEEE Transactions on Knowledge and Data Engineering, Vol 7, No. 4, August, 1995.