Spatio-Temporal Video Search Using the Object Based Video Representation

Di Zhong and Shih-Fu Chang

Image and Advanced Television Laboratory
Department of Electrical Engineering
Columbia University, New York, NY 10027, USA
{dzhong,sfchang}@ctr.columbia.edu

Abstract

Object-based video representation provides great promises for new search and editing functionalities. Feature regions in video sequences are automatically segmented, tracked, and grouped to form the basis for content-based video search and higher levels of abstraction. We present a new system for video object segmentation and tracking using feature fusion and region grouping. We also present efficient techniques for spatio-temporal video query based on the automatically segmented video objects.

1. Introduction

New functionality of video applications, such as search and editing, require a flexible framework of video representation, which allows for content based access. We have proposed a hierarchical object-based video model for this purpose [8]. It facilitates efficient search at the low level of features as well as the high level of semantic abstraction. This model is also synergistic with the new video compression standard such as MPEG-4 [2].

The hierarchical object-based model consists of image regions with consistent features at the bottom level, the video objects satisfying spatio-temporal constraints at the intermediate level, and the semantic abstraction classes at the top level. To build an efficient video indexing schema using the object model, we need to address two main issues. First, salient feature regions need to be extracted and grouped to form the basis for higher levels of abstraction and search functionality. Secondly, efficient techniques for indexing and searching based on the spatio-temporal structures of video objects need to be developed.

Automatic segmentation and tracking of video object is a difficult process. The goal is to segment and track regions with consistent visual properties, such as color, texture, motion, or spatio-temporal structures. To develop a system that works well for general videos, fusion of variant visual features in the segmentation process is essential. In [3], boundary of motion segmentation is enhanced with color segmentation. A modified watershed algorithm is developed in [4] to segment the combination of spatial and temporal gradient image. In [5], color based segmentations are fused with edge images to segment image objects. B-spline based matching algorithms are proposed to match similar object shapes in the database.

In the area of content based video indexing and query, much progress has been made in recent years, such as object searching based on localized color [11], moving object extraction and indexing [10], motion trail matching in MPEG [14], key-frame based video browsing [12], and abstraction of high level video units [9]. Although feature based indexing techniques have been proved to be useful, it's still difficult to link the low-level features to semantic meanings. To partly alleviate this problem, visual search according to the spatio-temporal structures of video objects allows users to find video from a large archive by describing the visual scenes as much as possible using powerful graphic tools.

Based on symbolic representation of images, some techniques such as 2D-strings [1] and spatial quad-trees [13] have been utilized for flexible and efficient spatial query. In [6], 2D-strings is extended with temporal logics to handle temporal structures. Visual queries using integrated spatial and feature based indexing has been demonstrated in [7] recently.

In this paper, we first present an improved system, based on [8], for video object segmentation and tracking using feature fusion and region grouping. In section 3, we discuss the construction of visual feature library and present efficient techniques for spatio-temporal query of video objects. Experimental results from a new prototype for video searching, VideoQ, is given in section 4.

2. Video object segmentation using feature fusion and region grouping

The segmentation and tracking of feature regions is based on the fusion of color, edge and optical flow. Color is chosen as the major segmentation feature

because of its consistency under varying conditions. As boundaries of color regions may not be accurate due to noises and image filtering process that is usually applied in the color segmentation stage, edge information is incorporated into the segmentation process to improve the accuracy. Optical flow is utilized to project and track color regions through a video sequence. Motion feature can also be used to group multiple feature regions to form higher-level video objects, such as people or vehicles.

The basic region segmentation and tracking procedure is shown in **figure 1**. Projection and segmentation is the major module in which different features are fused. This module is further described in **figure 2**.

Optical flow of current frame n is derived from frame n and n+1 in the motion estimation module using a hierarchical block matching method. Taken color regions and optical flow generated from above two processes, a linear regression algorithm is used to estimate the affine motion for each region. Affine motion parameters are further refined by using a log-step region matching method in the six-dimensional affine space. Through above modules, color regions with affine motion parameters are generated for frame n. Similarly, these regions will be tracked in the segmentation process of frame n+1. A region grouping module is applied at the final stage to avoid over-segmentation and obtain higher-level video objects.

In the projection and segmentation module (**figure 2)**, at the first step, the current frame(n) is quantized in a perceptually uniform color space (e.g., CIE L*u*v* or HSV) and smoothed with a non-linear median filter. Its edge map is also extracted. Then an interframe projection algorithm is used to track segmented regions from the previous frame (frame n-1). Existing regions in frame n-1 are firstly projected into frame n according to their motion estimations (i.e. affine parameters). For each pixel in frame n, if it is covered by a projected region and the color difference between the pixel and the mean color of the region is below a given threshold, the pixel is labelled as the old region. Outstanding pixels remain un-labelled at this point. In the next step, these tracked regions together with un-labelled pixels are further processed by an intraframe segmentation algorithm. An iterative clustering algorithm is adopted: two adjoining regions with the smallest color distance are continuously merged until the difference is larger than a given threshold. Overall, the procedure generates homogeneous regions in frame n based on tracking of existing regions from frame n-1.

In both the inter- and intra-segmentation processes mentioned above, only non-edge pixels are processed and labelled. Edge pixels are not merged into any regions in these processes. Regions clearly separated by long edge lines will not be spatially connected and thus will not be merged. After the merging of all non-edge pixels, edge pixels are assigned to their neighboring regions according to color similarity. We have found this greatly enhances the accuracy of region boundaries.

Compared with our previous results [8], the new segmentation experiment with Foreman sequence (QCIF) shows obvious improvement (**Figure 3).** The top row shows original sequence. The second row shows two major tracked regions out of about 10 segmented regions. Regions are shown with their representative (i.e. average) colors. Interesting regions and events (such as eyes open, mouth opens) can be detected as

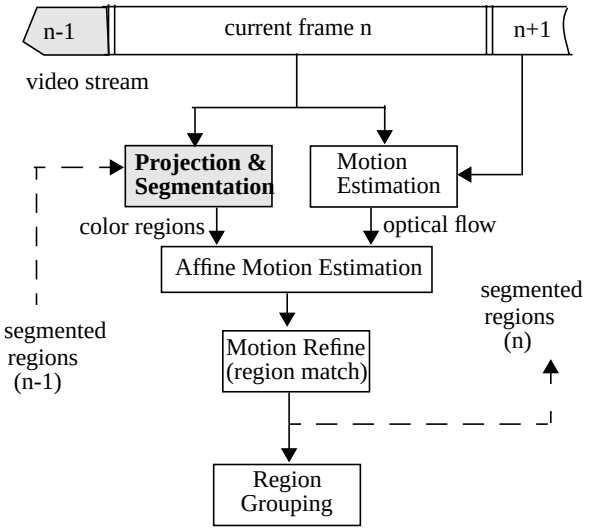

Figure 1. Video object segmentation and tracking

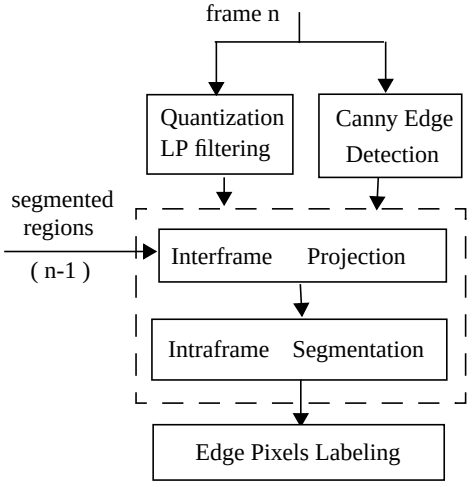

Figure 2. Projection and segmentation module (the shaded block in Fig. 1)

well. Another experiment with a baseball sequence is also included. Major regions of the player are clearly segmented and tracked over time. Experiments show that our system is robust for the tracking of salient color regions under different circumstances, such as multiple objects, and fast/slow motion.

To group regions to objects at a higher level, several criteria are adopted. First, adjacent regions with similar motions are grouped into one moving object. This is realized by using affine motion parameters of each region. Sizes and durations are utilized to eliminate noisy and unimportant regions: regions with small size and small duration are merged with its neighbor regions. Temporal evolution of regions are also considered in the grouping process to generate consistent region groups. Furthermore, model based grouping can also be fulfilled using the spatio-temporal matching techniques that will be discussed in section 3.

3. A framework for spatio-temporal video search

To build a video object database for content search, long video streams are first segmented into "shots" using scene cut detection algorithms. Then the segmentation and tracking processes are used to extract salient regions from each shot. Different visual features such as color, texture, shape and trajectory of tracked regions are computed and stored in the visual feature library. These features are used as a foundation for content based searching according to spatio-temporal queries and high-level abstraction of semantic objects.

Generalization of 2D-strings to similarity-based image region matching has been successfully demonstrated in the VisualSEEk system [7]. It provides a framework for searching for and comparing images according to the spatial arrangement of automatically segmented feature regions. Multiple regions can be queried either by their absolute or relative locations. The overall query strategy consists of individual matching of each local region and a "join" operation of the query results from matching individual regions. The query process identifies the candidate target images which contain matches to all query regions, and then verifies the absolute or relative spatial constraints to determine the best images that satisfies both the visual features and the spatial relationships among regions.

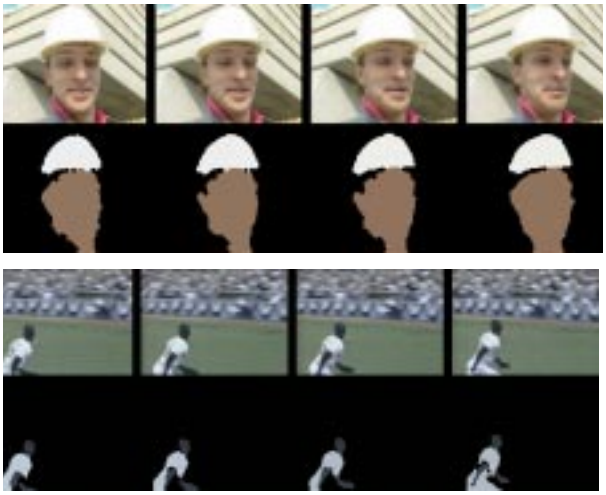

Figure 3. Region segmentation using feature fusion

There are several options to extend the above model to index the spatio-temporal structures in video. The first approach only indexes frames with significant changes of spatial structures. For example, a 2D-string followed by a sequence of edits can be applied [15]. Given such a representation, users will be able to search video objects or events of interest (e.g., objects with a specific appearance order, objects changing positions) by specifying changes of spatial structures over time.

Combination of motion trajectory matching with 2D strings is another approach. For a set of query objects, we first match their individual trajectories. This results in a list of candidate video objects for each query object. Then, spatial constraints are verified at selected sampling frames using 2D strings to verify spatial relationships among video objects at these time points. As trajectory matching returns video objects with similar motion trajectories, examining the spatial structure at selected frames is sufficient to approximately validate the entire spatio-temporal structure.

4. Experimental results

Our experimental setup includes 200 video shots, covering various types of content, such as sports, nature, science, and history. By applying the object segmentation and tracking algorithm, we generate more than 2000 salient video objects. Currently, for each video object, the following visual features are indexed: representative color, Tamura texture, aspect ratio, normalized size, and motion trajectory. Currently, the trajectory is represented as the coordinate sequence of the object centroid at successive frames after global motion compensation.

The Web-based interface of the VideoQ system (http://www.ctr.columbia.edu/VideoQ) is shown in **figure 4**. Users can specify multiple motion objects with different visual features. When a query is

submitted to the server, the searching engine will find video shots with objects satisfying the spatio-temporal structures specified in the query, and return the icons of these shots to users. Users may click on any icon to play the video shot at the full resolution and full rate.

We have conducted performance tests using sample queries and several performance metrics. One metric is to measure the precision/recall for each sample query. But precision/recall measurements require manual assignment of relevance of each video in the database to the query input and thus is impractical. Another metrics will to measure the overall time for the user to find specific video clips of interest by using the searching tools provided. We have found that through multiple iterations of searches and query refinement, users are usually able to find right video clips within few iterations and within a short time.

In real applications, we have found that this type of tools are very appealing for general users. However, for users in the professional practice (e.g., video studios), this new way of searching is different from what they have been used to. The new paradigm of object-oriented content-based video search presented in this paper is interesting, but its real impact still needs to be carefully examined in actual applications.

5. Conclusions

We describe a new video object segmentation and tracking system using feature fusion and region grouping. Based on this object-based video representation framework, we develop a powerful content-based video search system which indexes the spatio-temporal structures of the automatically segmented video objects. Promising experimental results from our research prototype, VideoQ, are presented. Evaluation in large scale applications is necessary to examine the impact of this new generation of video search systems on various types of users.

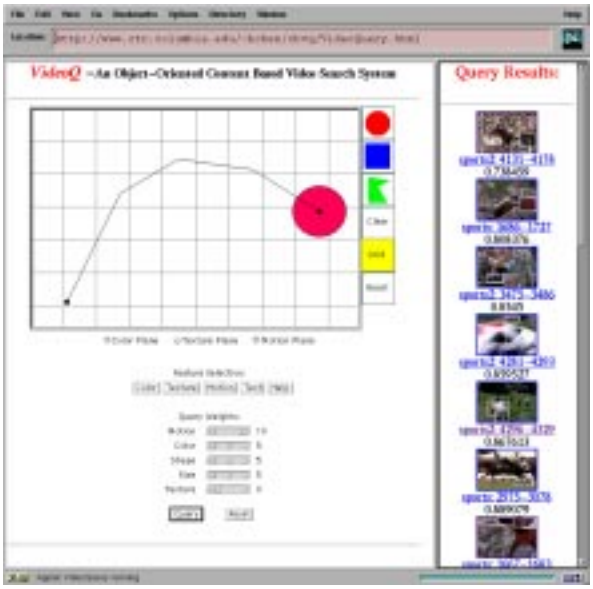

Figure 4. The Web-based user interface of VideoQ

6. References

[1] 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

[2] ISO/IEC ISO/IEC JTC1/SC29/WG11 N1797, "MPEG-4 Visual Working Draft Version 4.0", July 1997.

[3] C. Gu, T. Ebrahimi, M. Kunt, “Morphological Moving Object Segmentation and Tracking for Content-based Video Coding”, Multimedia Communication and Video Coding, New York, 1996

[4] J.G. Choi, Y.-K. Lim, M.H. Lee, G. Bang, J. Yoo, "Automatic segmentation based on spatio-temporal information", ISO IEC JTC1/SC29/W11 MPEG95/ M1284, Sept 1996

[5] E.Saber, A.M. Takalp & G. Bozdagi,"Fusion of color and edge information for improved segmentation and edge linking", IEEE ICASSP’96, Atlanta, GA, May 1996.

[6] 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, v7 August, 1995

[7] J.R. Smith and S.-F. Chang, "VisualSEEk: a fully automated content-based image query system", ACM Multimedia 96, Boston, MA, Nov 20 1996.

[8] D. Zhong and S.-F. Chang, "Video Object Model and Segmentation for Content-Based Video Indexing", ISCAS’97, HongKong, June 9-12, 1997

[9] M. M. Yeung, B. L. Yeo and B. Liu, "Extracting Story Units from Long Programs for Video Browsing and Navigation", International Conference on Multimedia Computing and Systems, June 1996.

[10] S.Y. Lee and H. M. Kao, "Video Indexing- An approach based on moving object and track", Proc. IS&T/SPIE, Vol. 1908, 1993, pp25-36.

[11] A. Nagasaka and Y. Tanaka,"Automatic Video Indexing and Full-Video Search for Object Appearances", Visual Database Systems II, 1992

[12] D. Zhong, H. J. Zhang and S.-F. Chang, "Clustering methods for video browsing and annotation", Storage & Retrieval for Still Image and Video Databases IV, IS&T/ SPIE's Electronic Imaging, Feb. 96

[13] H. Samet. The quadtree and related hierarchical data structures. ACM Computing Surveys, 16(2), 1984.

[14] N. Dimitrova and F. Golshani, "RX for Semantic Video Database Retrieval," ACM Multimedia Conference, San FRancisco, Oct. 1994.

[15] K. Shearer, S. Venkatesh, and D. Kieronska, "Spatial Indexing for Video Databases", Journal of Visual Communication and Image Representation, Volume 8, Number 3, September 1997.