To appear in IEEE Intern. Conf. on Circuits and Systems, June, 1997, Hong Kong.
(special session on Networked Multimedia Technology & Application)

Video Object Model and Segmentation for Content-Based Video Indexing

D. Zhong and S.-F. Chang

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

Abstract

Object segmentation and tracking is a key component for new generation of digital video representation, transmission and manipulations. Example applications include content-based video database and video editing. In this paper, we present a general schema for video object modeling, which incorporates low-level visual features and hierarchical grouping. The schema provides a general framework for video object extraction, indexing, and classification. In addition, we present new video segmentation and tracking algorithms based on salient color and affine motion features. Color feature is used for intra-frame segmentation; affine motion is used for tracking image segments over time. Experimental evaluation results using several test video streams are included.

1. Introduction

Object segmentation and tracking plays an important role for new generation of digital video compression, transmission and manipulations. Example applications include content-based video indexing, search and authoring [1,2]. By video objects, here we refer to objects of interest including salient low-level image regions (uniform color/texture regions), moving foreground objects, group of primitive objects satisfying spatio-temporal constraints (e.g., different regions of a car or a person). Automatic extraction of video objects at different levels can be used to generate a library of video data units, from which various functionalities can be developed. For example, video objects can be searched according to their visual features, including spatio-temporal attributes. High-level semantic concepts can be associated with groups of low-level objects through the use of domain knowledge or user interaction. Representing image sequences in terms of objects also has great synergy with the new generation video compression techniques such as those discussed in MPEG-4 [3,4].

In the area of video indexing and retrieval, it is recognized that text based indexing often is insufficient because of the time-consuming manual annotation process, biased subjective interpretation, and possible semantic ambiguity. Visual feature based indexing techniques have been proved to be useful and provide great complement. However, visual features tend to provide little direct links to high-level semantic concepts which are directly useful to users. One promising approach to overcoming this deficiency is to develop a hierarchical object representation model, in which video objects in different levels can be indexed, searched, and grouped to high-level concepts (e.g., a person, a car, a sports scene).

Video object segmentation is a difficult task. The goal is to segment image into regions with salient homogeneous properties, such as color, texture, motion, or spatio-temporal structures. In [5], morphological segmentation algorithms are used for intraframe and interframe segmentation of coherent motion regions from successive motion vector fields. To obtain accurate motion boundary, color based spatial segmentations are used to refine the motion segmentation results. In [6], a modified watershed algorithm is developed to segment the combination of spatial and temporal gradient image. The segmented regions are further merged based on the connection graph to obtain meaningful partitions. In [7], 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.

Generally, video segmentation and tracking are based on selective features and their consistence properties over space and time. However, due to the complexity of objects and possible occlusion, it is usually hard to find such consistent features directly for high-level objects. In this paper, we present a schema incorporating visual features for general video object representation, extraction and indexing. In section 2, we first describe the proposed hierarchical object model and its bottom-up construction process. Then in section 3, a new object segmentation and tracking system using salient color and optical flow is described. Experimental results and conclusions are given in section 4.

2. A Hierarchical Video Object Model for Video Indexing

As mentioned above, in general, it is hard to track a meaningful object (e.g., a person) due to its dynamic complexity and ambiguity over space and time. Objects usually do not correspond to simple partitions based on single features like color or motion. Furthermore, definition of high-level objects tends to be domain dependent.

On the other hand, objects can usually be divided into several spatial homogeneous regions according to image features. These features are relatively stable for each region over time. For example, color is a good candidate for low-level region tracking. It does not change significantly under varying image conditions, such as change in orientation, shift of view, partial occlusion or change of shape. Some texture features like coarseness and contrast also have nice invariance properties. Thus, homogenous color or texture regions are suitable candidates for primitive region segmentation. Further grouping of objects and semantic abstraction can be developed based on these basic feature regions and their spatio-temporal relationship. Based on these observations, we proposed the following model for video object tracking and indexing (**figure 1**).

At the bottom level are primitive regions segmented according to color, texture, or motion measures. As these regions are tracked over time, temporal attributes such as trajectory, motion pattern, and life span can be obtained.

Although homogeneous regions are potential candidates for tracking, usually the number of primitive regions is too large (e.g. 20-40). To limit the number of regions, multiple regions can be grouped based on spatial connectivity or temporal motion homogeneity. The motion features we refer to here include both instantaneous motion vectors at every frame (2D or 3D models) and long term trends. Use of long-term trends of motion pattern in region grouping may preclude it from real-time applications. However, adaptive decision can be made online based on evolution of region motion patterns.

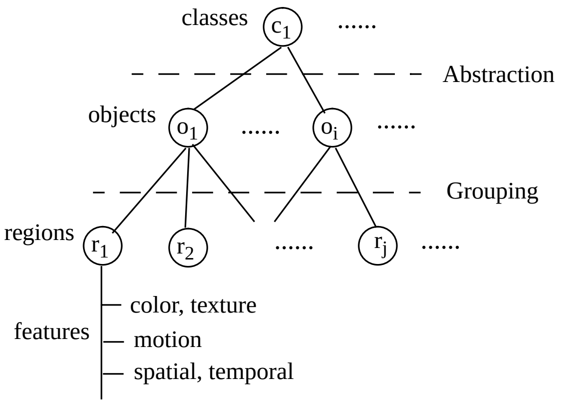

Figure 1: Hierarchical representation of video objects

The top level includes links to conceptual abstraction of video objects. For example, a group of video objects may be classified to moving human figure by identifying color regions (skin tone), spatial relationships (geometrical symmetry in the human models), and motion pattern of component regions. Usually, accuracy of automatic methods decreases as the level goes up. Fully automatic mapping of video objects to semantic concepts for unconstrained domains is still difficult. In order to solve this problem, several general approaches are taken in recent works. First, some minimal level of user input is used to label example video objects. The system then propagate these user-assigned labels to other video objects in the repository based on visual similarity. The second approach applied unsupervised or supervised clustering of video objects based on their visual features and then tries to map the clusters to subjective concepts [8]. In [9], the system learns from the positive or negative examples from the users and adapt the mapping between object clusters and subjective classes. Third, information in other media channels associated with the same video objects can be used in subject classification. In [10], the text information associated with the images/videos in online multimedia documents is used to achieve initial subject classification. Visual features are used for automatic image type classification (e.g., graphics vs. photographs, portrait vs. non-portrait images). Finally, other approaches achieve higher accuracy by constraining the systems to specific application domains and thus benefit from the use of domain knowledge (e.g., news video, sports video).

We propose the above hierarchical video object schema for content-based video indexing. One challenging issue here is to maximize the extent of useful information obtained from automatic image analysis tasks. A library of low-level regions and mid-level video objects can be constructed to be used in high-level semantic concept mapping. This general schema can be adapted to different specific domains efficiently and achieve higher performance.

3. Region Segmentation and Tracking using Color and Optical Flow

In this section, we describe a region segmentation and tracking system using color and optical flow. It generates primitive regions corresponding to the bottom level of the object schema in **figure 1**. We present this system to illustrate the feasibility of automatic segmentation of primitive regions and the potentials of video object model proposed in the previous section.

The basic region segmentation and tracking procedure is shown in **figure 2**. Its projection and segmentation module, where the main segmentation and tracking process takes place, is further described in **figure 3**. First, each image frame is quantized in a perceptually uniform color space, e.g., the L*u*v* space. Quantization templates can be obtained by the uniform quantizer or clustering algorithms (e.g., self-organization map). After quantization, non-linear median filtering is used to eliminate insignificant parts and outliers in the image while preserving edge information.

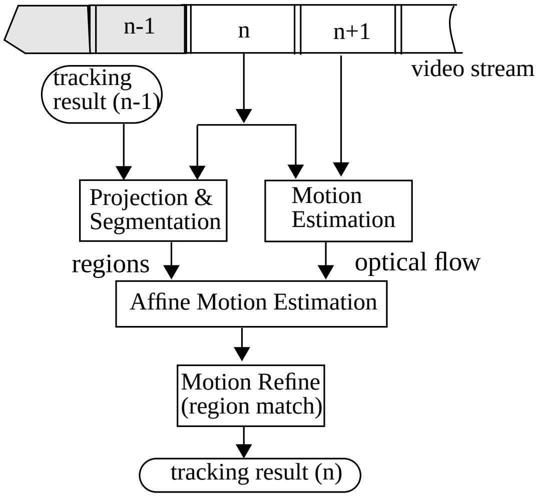

Figure 2. Region segmentation and tracking of frame n

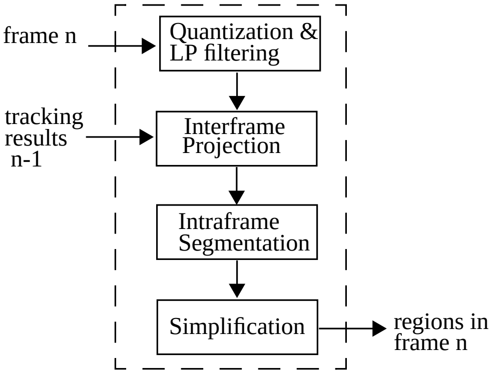

Figure 3. Region projection and segmentation

For the first frame in the sequence, the system uses only intraframe segmentation. For intermediate frames, as segmented regions with image features and motion information are available from frame n-1, an interframe projection algorithm is used to track regions. Conceptually, all existing regions at frame n-1 are first projected into frame n according to their motion estimations (affine parameters, see next paragraph). For every pixel in frame n, if it is covered by a projected region and the color difference (weighted Euclidean distance in L*u*v* space) between the pixel and the mean color of the region is under a given threshold, it is labelled as the same region. When there are more than one regions satisfying the above condition, the one with the smallest color difference will be chosen. If no region satisfies the condition, the pixel will remain un-labelled at this point.

The tracked regions together with un-labelled pixels are further processed by an intraframe segmentation algorithm. New uncovered regions will be detected in this process. An iterative clustering algorithm is adopted: two adjoining regions with the smallest color distance are continuously merged until the difference is larger than the given threshold. Finally, small regions are merged to their neighbors by a morphological open-close algorithm. Thus, the whole procedure generates homogeneous color regions in frame n while tracking existing regions from frame n-1. As one can see, the distance measure in the projection and segmentation algorithms can be easily changed or expanded to include other features like texture.

Other modules in **figure 2** are described as follows. Optical flow of frame n is derived from frame n and n+1. A hierarchical block matching method is adopted to compute the displacement vectors of each pixel. Different from simple block matching methods which estimate motion solely based on minimum mean square errors, this technique yields reliable and homogeneous displacement vector fields, which are close to the true displacements [11]. Then a linear regression algorithm is used to estimate the affine motion of each region from known displacements of pixels inside the region. The affine transformation is a good first-order approximation to a distant object undergoing 3D translation and linear deformation. Finally, the affine motion parameters are further refined by using a log(D)-steps region matching method. It is similar to usual block matching, while the difference is that the search is done in the six-dimensional affine space.

In short, the tracking results of frame n include a list of regions with color features, motion parameters, as well as motion trajectories. All these will be passed to the system again for segmentation of frame n+1.

4. Experimental Results and Conclusion

The region segmentation and tracking results of two sequences are given in **Figure 4.** In each case, the top row shows original sequence. The second row shows a selected subset of automatically segmented regions being tracked. Tracked regions are all drawn using their representative (i.e. average) colors. In both sequences, major regions are well segmented and tracked through the whole sequence. Experiments show that our system is robust for the tracking of salient color regions under different circumstances, such as multiple objects, fast/slow motion and region covering/uncovering.

In the Foreman sequence, by further considering the similarity of motions between the cap and the head, we can easily group them into one moving object at the higher level. The missing holes in the face region (not shown) are also tracked well and can be grouped to the head object based on motion similarity. Similarly to intraframe segmentation, an iterative contiguous clustering algorithm can be applied for grouping. For the baseball sequence, the motions of different parts of the player are slightly different. However, considering trajectories and spatial relationship of these regions in the long term, it is also possible to group them to one single moving object.

Our current work focuses on enhancement using other visual features such as texture. We also focus on techniques for grouping primitive regions to intermediate objects based on their spatio-temporal structures. Using the general video model schema proposed in section 2, we are studying learning and clustering techniques for extraction of high level objects and classes for semantic-level video annotation and indexing.

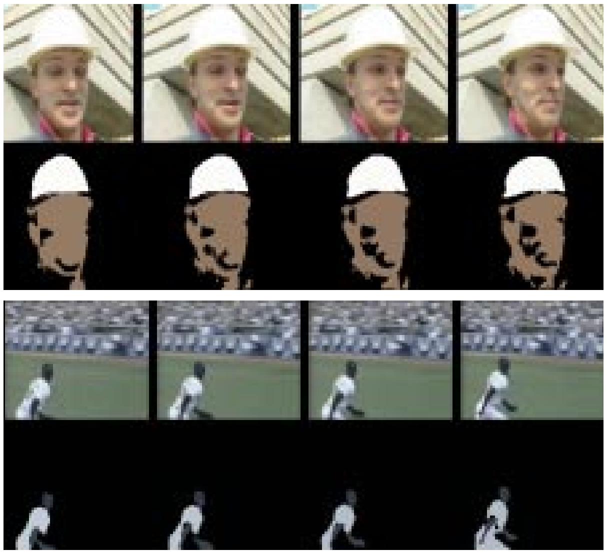

Fig. 4 Two test examples of region segmentation & tracking.

References

[1] H. S. Sawhney, "Motion Video Annotation and Analysis - An Overview", Conference Record of the 27th ASILOMAR Conf. on Signals, Systems & Computers, Vols. 1 & 2 ed. by A. Singh. IEEE Computer Society Press, 1993. pp. 85-89

[2] J. Meng and S.-F. Chang, "CVEPS: A Compressed Video Editing and Parsing System," ACM Multimedia Conference, Boston, MA, Nov. 1996. (demo: *http://www.ctr.columbia.edu/cveps*).

[3] "Description of MPEG-4", ISO/IEC JTC1/SC29/ WG11 N1410, MPEG document N1410 Oct. 1996.

[4] C.-T. Chu, D. Anastassiou and S.-F. Chang, “Hybrid Object-Based/Block-Based Coding in Video Compression at Very Low Bitrate”, to appear in J. of Image Communication-Signal Processing, Special Issue on MPEG-4, 1997.

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

[6] 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

[7] E. Saber, A.M. Takalp, & G. Bozdagi, "Fusion of Color and Edge Information for Improved Segmentation and Edge Linking," in IEEE ICASSP 96, Atlanta, GA, May 1996.

[8] 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: Science & Technology, Feb. 96

[9] T. P. Minka and R. W. Picard, "Interactive Learning using a Society of Models", MIT Media Lab Perceptual Computing TR #349, 1996.

[10] J. R. Smith and S.-F. Chang, "Searching for Images and Videos on the World-Wide Web," submitted to IEEE Multimedia Magazine, 1996. (also CU/CTR Technical Report #459-96-25,
demo: http://www.ctr.columbia.edu/webseek).

[11] M. Bierling, “Displacement Estimation by Hierarchical Block Matching”, SPIE Vol 1001 Visual Communication & Image Processing, 1988.