AUTOMATIC VIDEO SEGMENTATION AND INDEXING – WHERE ARE WE?

A. Murat Tekalp

Koc University, College of Engineering
Rumelifeneri Yolu, 34450 Sariyer, Istanbul, Turkey

1. INTRODUCTION

Video segmentation includes: 1) Temporal segmentation, such as shot boundary detection and special effects detection; and 2) Spatio-temporal segmentation, such as video object segmentation and tracking.

Video indexing includes: 1) Low-level indexing, such as using visual features such as color, texture, shape and motion; 2) Semantic level indexing, such as shot classification, story segmentation, people detection and recognition, etc.; and 3) Video summarization, such as key frame summarization, important event detection, etc.

Clearly, the accuracy of automatic video indexing strongly depends on the accuracy of automatic video analysis, and in particular automatic video segmentation. In the following, we first discuss where we are in automatic video segmentation, so that we can assess where we are in automatic video indexing better.

2. VIDEO SEGMENTATION

Video segmentation includes temporal segmentation and spatio-temporal segmentation.

1) Temporal Segmentation: Temporal segmentation is relatively an easier problem compared to spatio-temporal video object segmentation and tracking. Shot boundary detection methods locate frames, across which large differences are observed in some feature space [1-3]. The feature space usually consists of a combination of color and motion. Boundaries can be sharp, known as cuts, or gradual, called special effects, such as wipes and fades. It is easier to detect cuts than special effects. The simplest method for cut detection is to check pixel intensity differences between successive frames. If a pre-determined number of pixels exhibit differences larger than a threshold value, then a "cut" can be declared. A slightly different approach may be to divide each frame into rectangular blocks, employ statistical tests within each block independently, and then check the count of changed blocks against a set threshold. Both approaches can be sensitive to presence of noise and compression artifacts in the video. However, solutions that apply to generic video with reasonable accuracy exist [1-3].

2) Spatio-temporal Segmentation: Video object segmentation is not an easy problem, mainly because definition of video objects usually requires semantic level interpretation of the scene. It is generally not possible to define such semantically meaningful video objects in terms of coherent low level features, such as uniform color or uniform motion parameters.

It is possible to automatically segment coherent spatio-temporal regions based on low-level features, such as planar objects with rigid motion or regions with uniform color, using techniques such as K-means clustering or Bayesian methods. For example, NeTra-V extracts and tracks spatio-temporal regions within shots for object-based indexing and retrieval [4]. Similarly, VideoQ employs color and edge information for segmentation of regions and motion information to track them [5]. The segmentation is re-initialized at each shot boundary; therefore, regions are not linked across shots. In general, these procedures result in over-segmentation of the scene. Hence, segmentation and tracking of semantic objects in an unconstrained scene may require interactive user intervention. However, in some well-constrained settings, semantic objects can be computed fully automatically. For example, in video surveillance systems [6,7], where the camera is stationary, objects in the scene can be extracted by simple change detection and background subtraction methods.

Once correctly initialized, video objects can be tracked with some degree of accuracy in most cases, except when there exist significant self-occlusions or multiple moving objects with unpredictably intersecting trajectories. Object tracking problems can be categorized as i) trajectory tracking and ii) boundary tracking problems. The former problem where we need to track a few control points, such as the center of mass or the bounding box, is relatively easier. However, some applications, such as video synthesis and editing, require pixel-accurate tracking of the exact object boundary, which is a significantly more complex and difficult task.

3. VIDEO INDEXING AND SUMMARIZATION

Video indexing includes low-level indexing, semantic level indexing, and video summarization [8,9].

1) Low-level Indexing: Low-level descriptors, such as color, texture, shape, and motion, can be associated with shots or objects. Color of selected keyframes or all frames in a shot can be described by color histogram or dominant color descriptors [10]. Camera motion and motion activity parameters describe shot-level motion [10,11]. Object motion can be described by motion trajectory [12].

2) Semantic-level Indexing:
Semantic information can be represented by structured or free-text annotations, or by semantic models. Annotations can be manual, or extracted automatically from closed-captions, on-screen text, or by face detection and recognition. Semantic models can support entities, such as objects and events, and relations between them, which make processing complex queries possible. Semantic video models can be considered as extensions of entity-relation (ER) models developed for documents by the database and information retrieval communities.

3) Video Summarization:
Keyframe summaries and important segment summaries are commonly used in commercial applications. Keyframes, which refer to one or more representative frames in a shot, provide a compact visual representation. Several methods exist to automatically select keyframes by low-level feature analysis [8,9].
Semantic analysis of video generally involves use of both cinematic and object-based features. Cinematic features, that originate from common video composition and production rules, such as shot types and transitions. Different cinematic rules may apply to different domains of content. For example, action movies, TV sit-coms, TV news, and sports broadcasts all have different cinematic features [13-15]. Object segmentation and recognition methods can also be used for important segment/event detection [16].

4. CONCLUSION

While successful results can be obtained for certain video analysis tasks such as shot boundary detection on generic video, most automatic analysis tasks work better in specific content domains. Hence, video indexing and summarization at the semantic level can achieve success only in specific content domains.

5. REFERENCES

1) U. Gargi, R. Kasturi, and S. H. Strayer, "Performance characterization of video-shot change detection methods," IEEE Trans. Circ. Syst. for Video Tech., vol. 10, pp. 1-13, Feb. 2000.
2) R. Lienhart, "Reliable transition detection in videos: A survey and practitioner's guide," Int. J. Image Graph., vol. 1, pp. 469-486, Aug. 2001.
3) A. Hanjalic, "Shot-boundary detection: Unraveled and resolved?" IEEE Trans. Circ. Syst. for Video Tech., vol. 12, pp. 90-105, Feb. 2002.
4) Y. Deng and B. S. Manjunath, "NeTra-V: Toward and object-based video representation," IEEE Trans. Circ. Syst. for Video Tech., vol. 8, no. 5, pp. 616-627, 1998.
5) D. Zhong and S. F. Chang, "An integrated approach for content-based video object segmentation and retrieval," IEEE Trans. Circ. Syst. for Video Tech., vol. 9, no. 8, pp. 1259-1268, Dec. 1999.
6) J. D. Courtney, Automatic video indexing via object motion analysis, Pattern Recognition, vol. 30, no. 4, pp. 607-626, April 1997.
7) G. L. Foresti, L. Marcenaro, and C. S. Regazzoni, "Automatic detection and indexing of video-event shots for surveillance applications," IEEE Trans. Multimedia, vol. 4, no. 4, pp. 459-471, Dec. 2002.
8) N. Dimitrova, H. Zhang, B. Shahraray, I. Sezan, T, Huang, and A. Zakhor, "Applications of video content analysis and retrieval," IEEE Multimedia, vol. 9, pp. 42-55, July-Sept. 2002.
9) S. Antani, R. Kasturi, and R. Jain, "A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video," Pattern Recognition, vol. 35, pp. 945-965, 2002.
10) B. S. Manjunath, P. Salembier, and T. Sikora (Eds), Introduction to MPEG-7: Multimedia Content Description Interface, Wiley, 2002.
11) Y. P. Tan, S. R. Kulkarni, and P. J. Ramadge, "Rapid estimation of camera motion from compressed video with application to video annotation," IEEE Trans. Circ. Syst. Video Tech., v. 10, pp. 133-146, 2000.
12) S. Dagtas, W. Al-Khatip, A. Ghafoor, and R. L. Kashyap, "Models for motion-based video indexing and retrieval," IEEE Trans. Image Proc., vol. 9, no. 1, pp. 88-101, Jan. 2000.
13) A. Hampapur, R. Jain, T. E. Weymouth, "Production model based digital video segmentation," Multimedia Tools Appl., vol. 1, pp. 9-46, 1995.
14) H. Sundaram and S.-F. Chang, "Computable scenes and structures in films," IEEE Trans. Multimedia, vol. 4, pp. 482-491, Dec. 2002.
15) A. Ekin, A. M. Tekalp, R. Mehrotra,, "Automatic soccer video analysis and summarization," IEEE Trans. on Image Proc., vol. 12, no. 7, pp. 796-807, July 2003.
16) S. Satoh, Y. Nakamura, and T. Kanade, "Name-It: Naming and detecting faces in news videos," IEEE Multimedia, vol. 6, no. 1, pp. 22-35, Jan.-Mar. 2001.