STRUCTURAL AND SEMANTIC ANALYSIS OF VIDEO

Shih-Fu Chang *Hari Sundaram*
Department of Electrical Engineering
Columbia University
New York, NY 10027, USA
Email: {sfchang, sundaram}@ctr.columbia.edu

ABSTRACT

In this paper we discuss our recent research and open issues in structural and semantic analysis of digital videos. Specifically, we focus on segmentation, summarization and classification of digital video. In each area, we also emphasize the importance of understanding domain-specific characteristics. In scene segmentation, we introduce the idea of a computable scene as a chunk of audio-visual data that exhibits long-term consistency with regard to several audio-visual properties. In summarization, we discuss shot and program level summaries. We describe classification schemes based on Bayesian networks, which model interaction of multiple classes at different levels using multi-media. We also discuss classification techniques that exploit domain-specific spatial structural constraints as well as temporal transitional models.

1. INTRODUCTION

In this paper, we describe our recent research in developing algorithms and tools for segmentation, summarization and classification of video data. Segmenting video data into shots is often the first step in video analysis. Robust scene segmentation is an important step towards semantic understanding of video data. Summarization schemes based on such shot and scene segmentation facilitate non-linear navigation of the data. Classification techniques based on audio-visual as well as textual descriptors will allow us to categorize video data automatically.

There has been much prior work in each of three areas. In [12] the authors use scene transition graphs to segment the video using image data alone, into scenes. Their method assumes the presence of structure within a scene. While this might be present within interviews and other news programs, it is often absent in many scenes. In [4] the authors expanded video scene segmentation algorithms by adopting a psychological memory-based model.

In [12], the authors present program summaries using images. These images are derived from centroids of time-constrained shot clusters. In [9], the authors generate summaries by using an automatic clustering technique on the frames of the entire video data. However, both these are *static* summarization schemes.

We derive scene segmentation algorithms based on the idea of a computable scene [7]. The computable scene model was derived using rules from film-making and from experimental observations in the psychology of audition. A computable scene exhibits long-term consistency with respect to three properties: (a) chromatic composition of the scene (b) lighting conditions and (c) ambient audio. We term such a scene as *computable*, since it can be reliably computed using low-level features. Central to the idea of a computable scene is the notion of a causal, finite memory listener model [7] [8]. Such computable scene models integrate both audio and video features in scene analysis. They also facilitate exploration of complementary properties of audio and visual scene boundaries.

For classification, we apply the notion of recurrent visual semantics [3] and develop an interactive learning system for generating scene classifiers. They work well on highly structured domains such as sports (e.g. baseball games). In structured domains (sports and medicine), we also explore the temporal transition constraints in recognizing individual scenes. In [6], we use probabilistic reasoning networks to classify images into generic (indoor/outdoor) as well as topical classes (e.g., politics, crime etc.). The network includes classes at multiple levels and classifiers utilizing multiple media.

The rest of this paper is organized as follows: In section 2 we discuss shot and scene level segmentation. In section 3, we describe summarization schemes. In section 4, we discuss video classification schemes. Finally in the last section, we present our conclusions.

2. SEGMENTATION

In this section we discuss techniques for shot and scene level segmentation. We also introduce the notion of computable scenes.

2.1 Shot Level Segmentation

A shot is a single continuous camera take at a physical location. Segmenting video data into shots is often the first step in video analysis. Some popular techniques use change of color distribution or frame-to-frame correlation in order to detect shot boundaries. In most cases, shot segmentation tools are quite robust and can achieve a high accuracy (higher than 95%).

There are several interesting issues that arise in some practical domains. In the consumer home videos, acute brightness changes caused by a flash or due to automatic lighting control mechanisms (e.g. a sudden, automatic aperture change can take place due to unbalanced ambient lighting) usually result in false alarms. False alarms can also occur when there is a rapid change in the foreground (e.g., a person walks into and out of the scene rapidly). In broadcast news or documentaries, changes in the graphic (e.g., statistics on the screen) are usually not adequately segmented due to the consistent background and subtle changes in the content.

Aligning the results of the shot boundary detection algorithm with other media (speech, text captions) proves to be a particularly challenging problem. In broadcast news, the text

captions usually lag behind the video shots by a few seconds. It is possible to correct such lags by aligning the closed captions to the corresponding speech signal. However, in some cases video shot boundaries can be out of sync with the audio boundaries. For example, a news story may consist of several short video shots while the commentary may be spread over all the shots. In this case, it is difficult to break the speech into complete sentences in order to align with the video shots.

2.2 Scene Level Segmentation

A semantic scene is defined as a collection of shots that are consistent with respect to a certain semantic. For example, a collection of shots taken with a handy-cam at the beach would correspond to a semantic scene. However, determining the semantic of a scene. In [7], we determine constraints on a computable scene model by using film-making rules and experimental results in the psychology of audition. We use these constraints along with five hours of commercial film data to come up with two broad categories of computable scenes:

1. **N-type:** These scenes are characterized by a long-term consistency of chromatic composition, lighting conditions and sound.
2. **M-type:** These (or montage/MTV scenes) are characterized by widely different visuals (differences in location, time of creation as well as lighting conditions) which create a unity of theme by manner in which they have been juxtaposed. However M-type scenes are assumed to be characterized by a long-term consistency in the audio track.

We term this scene model to be *computable*, since all the features can be reliably and automatically determined using low-level features present in the data. This model covers a subset of the different computable scenes[1]. Note that each scene is a continuous chunk of audio-visual data. Earlier approaches [4] to semantic scene segmentation actually work in a computable model framework. We investigate four particular kinds of scenes that belong to the two aforementioned categories: (a) progressive (b) dialogue (c) MTV and (d) transient. The first two are N-type scenes while the last two are M-type scenes. We now briefly describe each scene.

Progressive: There can be a linear progression of visuals without any repetitive structure (the first part of figure 1 is progressive). For example, consider the following scene: Alice enters the room looking for a book. We see the following shots (a) she enters the room (b) she examines her book-shelf (c) looks under the bed (d) locates the book and the camera follows her as she sits on the sofa to read.

Dialogue: A simple repetitive visual structure (amongst shots) can be present if the action in the scene is a dialogue. A repetitive structure is also present when the film-maker shuttles back and forth between two shots (e.g. man watching television). We denote this as a thematic dialogue [7]. The latter half of figure 1 shows a dialogue.

[1] In our pilot study, we found that about 75% of the scenes were N-type scenes.

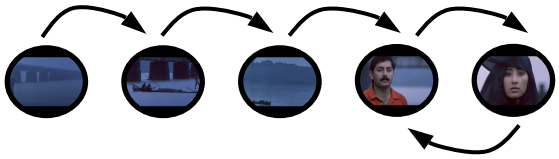

Figure 1: A progressive scene followed by a dialogue sequence.

In [7], we proposed a technique similar to a discrete periodic correlation to detect the alternating structure in a sequence of

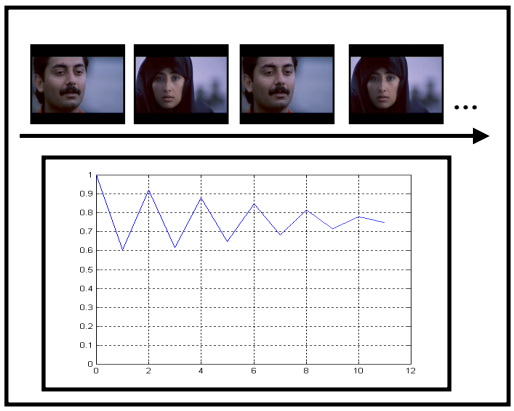

Figure 2: A dialogue scene in a commercial drama film and the graph of discrete periodic correlation. Note the distinct peaks at n=2, 4. . indicating a strong alternating structure.

shots. Preliminary results show an accuracy of about 95% based on a difficult test set: an hour each from five diverse commercial drama films. Figure 2 shows an example of a dialogue scene from a commercial drama film and the correlation sequence.

MTV: These scenes usually occur in music videos, TV commercials, or special scenes of films. Such videos typically have rapidly changing visual shots but consistent sound (usually the music).

Transient: Here dissimilar video shots progress at a slower pace than that in MTV scenes. Like MTV scenes, transient scenes have a consistent sound to convey a consistent semantic. Transient scenes usually occur when a thematic theme takes place over multiple physical locations.

The computable scene model implies that we need to use both audio and visual features for scene boundary detection. In [8], we demonstrate a causal, finite-memory based model for detecting both the audio and the video scene boundaries. There are several interesting issues that occur when combining audio and video scene boundaries. The audio and the video scene changes are usually synchronized but more interestingly, audio and video scenes can also change independently. These are important as they usually refer to thematic changes in the video. For example, the director can show a change in the mood of he film by simply changing the background audio. In transient scenes the video scene changes but the audio remains the same. Figure 3 illustrates this idea. Detecting scene changes in audio-visual streams is still a challenging issue, particularly for MTV and transient scenes.

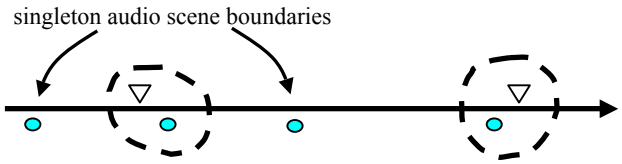

Figure 3: Complementary video (triangles) and audio (solid circles) scene change locations. The dashed circles show audio/video scene boundaries which agree.

3. SUMMARIZATION

We can create shot-level summaries by displaying shots as a time ordered sequence or by creating hierarchical clusters on key frames of the segmented shots [12]. Such summarization schemes have been used in several systems and have been included in the current draft of MPEG-7 standard [2]. There are several promising commercial solutions for presentation videos [9]. In [9], the system augments the key frames of each shot with key phrases extracted from the associated text transcript. It also provides a spatial mosaic interface as the summary of multiple shots.

We can also create program-level summaries. This can be done by determining the structural relationships amongst the constituent scenes in a video program. For example, similar visual scenes with different sounds may be indicative of the characters returning to an earlier physical location, while similar sounds with different visuals may represent scenes with a common theme (or mood such as joy or sadness etc.).

Topic detection and tracking using text-based techniques should prove to be very useful in generating program level summaries. Additionally, topic summarization of multiple videos is important when we deal with a large collection of videos [10].

Representing computable scenes such as dialog, MTV etc. is an interesting research problem. Additionally, current summarization techniques are static i.e. they do not allow the viewer to visualize or experience the dynamic structure extant in each scene.

Specific summarization strategies are possible in particular domains. For example, in interview or meeting videos, a summary of the participants is extremely useful. Users can then quickly see all the participants in the event, allowing them to randomly access the comments of each participant. Detection and recognition of people can be done by analyzing the video (i.e., face [11]), speech and text (including the transcript and embedded text). Integration of multimedia cues in summarization remains an interesting open issue.

4. CLASSIFICATION

We use classifiers to map video entities into one or more classes. These entities could be video objects, key frames, video shots, audio segments, scenes, or they could also be the entire video program. The classes can be generic, low-level categories (e.g. indoor/outdoor, people/no people, stationary/moving, city/landscape etc.) topical categories (e.g., politics, science, sports, etc) or specific event categories (e.g., the Yugoslavian crisis). Audio-visual features are more useful in recognizing the generic low-level classes while text-based features are more useful in recognizing the high-level classes.

Integration of multiple modes of data for improving classifier accuracy, is an interesting research problem. For example, videos of natural disasters usually show scenes of devastation, victims, or recovery efforts. Recognition of topical classes (e.g., through language processing of the captions or the embedded text) will help low-level classifiers of visual scenes. To provide an effective framework for integrating multiple classes and classifiers, we have developed a probabilistic reasoning system based on Bayesian networks, for news and consumer domains [6]. Such systems provide flexible inference capabilities in addition to significantly improving the classification accuracy. Figure 4 shows a Bayesian network for the domain of consumer photographs.

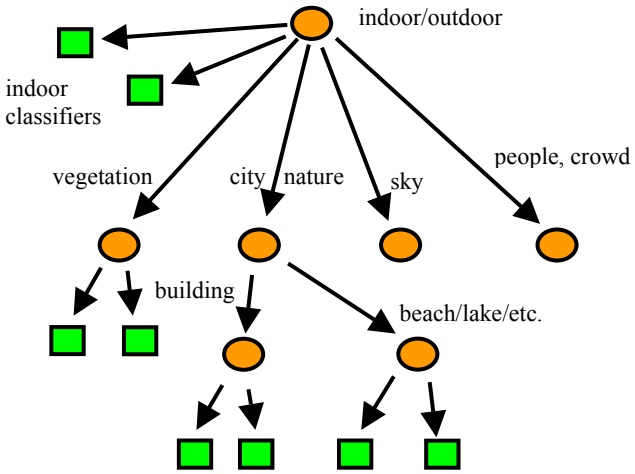

Figure 4: A Bayesian network for consumer photographs. It integrates multiple classes (circles) and classifiers (squares) using different media. Note the number of classifiers for each class can be arbitrary. Each link is associated with conditional probabilities modeling the dependence between the parent and the child class.

Integration of multimedia in decoding semantic classes of video content has also been explored in [5]. There the authors use a notion of a multijet to model events with strongly coupled audio-visual cues (e.g., explosions etc.). They also use the idea of a multinet, to model the interaction among different classes.

Specific domains can contain rich temporal transitional structures that help in the classification process. In sports, the events that unfold are governed by the rules of the sport hence contain a recurring temporal structure. The rules of production of such videos have also been standardized. For example, in baseball videos, there are only a few recurrent views, such as pitching, close up, home plate, crowd, etc. For medical videos, there is a fixed clinical procedure for capturing different video views. For example, in echocardiogram videos, there are a finite set of views (with distinctive structures and colors) corresponding to different transducer positions used in the clinical procedure.

In [3], we presented a new concept, that of recurrent visual semantics and an interactive learning system, *Visual Apprentice*. This system helps develop effective classifiers for recurrent views (e.g., pitching scene in a baseball game) in specific

domains. The learning system provides a flexible environment for users to define multi-level object-based models for arbitrary recurring scenes. the learning components present in the system help automatically determine the optimal set of features and classification algorithms for each part of the scene model. Figure 5 shows a multi-level scene model and a test image for the pitching scene.

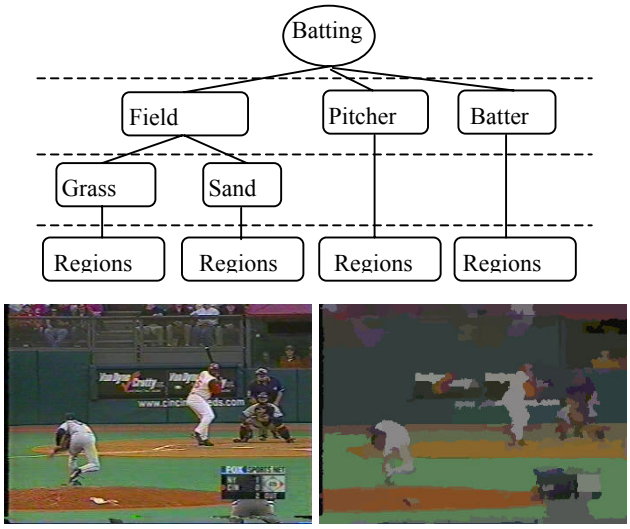

Figure 5. A multi-level scene model for the pitching scene in baseball. The bottom part shows a test baseball image and its segmented image map as the input to the multi-level classifier.

5. Conclusions

In this paper, we have presented a review of our research in segmentation, summarization and classification of audio-visual data. In each of the domains, we have laid emphasis on the importance of understanding characteristics of the content and the practical issues.

We discussed strategies for shot level segmentation and some issues (such as automatic aperture control and change in the foreground etc.) that make shot detection difficult in some practical situations. We introduced the idea of a computable scene as chunk of audio-visual data that can be automatically computed from low-level features in the data. The rationale behind this idea were film-making rules and experimental insights on the psychology of audition. We divided computable scenes into two broad categories N-type and M-type and described in some detail the properties of four particular scenes: (a) progressive (b) dialog (c) MTV and (d) transient.

We described existing shot level summaries and also discussed the idea of generating program level summaries by determining the relationships between the computable scenes in the program. We also discussed topic summarization and tracking using text based techniques.

We have shown that by using Bayesian networks we can integrate data from multiple media. Such systems can provide flexible inference systems as well as significantly improve the classifier accuracy. For specific domains that are governed by strict spatio-temporal rules, we have rich spatial and transitional structures that can be exploited in object/scene classification. We have applied this idea in developing an interactive classifier learning system, visual apprentice and by developing scene recognition algorithms in the sports and medical domains.

We also present interesting open issues in each area. In shot detection, the issue of dealing with sudden changes in the background as well sudden changes in lighting is an interesting problem. Aligning shot boundaries with other media is interesting. The computable scene model is limited in its scope and needs to broadened to cover more diverse scenes in a robust manner. The summarization schemes as noted earlier, suffer from being static in nature. Integration of multiple classes and classifiers at multiple levels using multimedia information is an interesting area. Finally, video scene recognition exploiting spatio-temporal structures in specific domains will prove fruitful in practice.

6. REFERENCES

[1] L. He et. al. *Auto-Summarization of Audio-Video Presentations,* ACM MM '99, Orlando FL, Nov. 1999.

[2] ISO/IEC JTC1/SC29/WG11 MPEG00/N3349, MPEG-7 Overview (Version 2.0), March 2000.

[3] A. Jaimes and S.-F. Chang, *Model Based Image Classification for Content-Based Retrieva*l, SPIE Conference on Storage and Retrieval for Image and Video Databases, Jan. 1999, San Jose, CA.

[4] J.R. Kender B.L. Yeo, *Video Scene Segmentation Via Continuous Video Coherence,* CVPR '98, Santa Barbara CA, Jun. 1998.

[5] M.R. Naphade, T. Kristjansson, B.J. Frey, and T.S. Huang, *Probabilistic Multimedia Objects (Multijects): A Novel Approach to Video Indexing and Retrieval in Multimedia Systems,* IEEE Intern. Conference on Image Processing, Oct. 1998, Chicago, IL.

[6] S. Paek and S.-F. Chang, *A Knowledge Engineering Approach for Image Classification Based on Probabilistic Reasoning Systems,* IEEE International Conference on Multimedia and Expo, New York, July, 2000.

[7] H. Sundaram and S.-F. Chang, *Determining Computable Scenes in Films and their Structures using Audio-Visual Memory Models,* Tech. Rep. 2000-05, also submitted to ACM Multimedia 2000.

[8] H. Sundaram and S.-F. Chang, *Video Scene Segmentation Using Video and Audio Features,* IEEE International Conference on Multimedia and Expo, New York, July, 2000.

[9] S. Uchihashi, J. Foote, A. Girgensohn, J. Boreczky, *Video Manga: Generating Semantically Meaningful Video Summaries,* ACM Multimedia '99, Orlando, FL, Nov. 1999.

[10] H. Wactlar, A. Hauptmann, Y. Gong, M. Christel, *Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library,* IEEE Computer 32(2): 66-73, 1999.

[11] H. Wang and S.-F. Chang, *FaceTrack- Tracking and Summarization Faces from Compressed Video,* SPIE Photonics East, Conference on Multimedia Storage and Archiving Systems, Boston, Nov. 1999.

[12] M. Yeung B.L. Yeo *Video Content Characterization and Compaction for Digital Library Applications,* Proc. SPIE '97, Storage and Retrieval of Image and Video Databases V, San Jose CA, Feb. 1997.