IEEE International Conference on Image Analysis and Processing (ICIAP), Montavo, Italy, Sept. 2003

Content-Based Video Summarization and Adaptation for Ubiquitous Media Access

Shih-Fu Chang

Dept. of Electrical Engineering, Columbia University
New York, NY 10027, USA
Ph.: + 1 212-749-5998 email: sfchang@ee.columbia.edu,
Web: http://www.ee.columia.edu/dvmm

ABSTRACT

Today's mobile and wireless users access multimedia content from different types of networks and terminals. Content analysis plays a critical role in developing effective solutions in meeting unique resource constraints and user preferences in such usage environments. Specifically, content analysis is central to automatic discovery of syntactic-level summaries and generation of concise semantic-level summaries. Content analysis also provides a promising direction for finding optimal adaptation methods under various resource-utility constraints. This paper presents brief overviews of such emerging, fruitful areas and promising research directions.

1. INTRODUCTION

Ubiquitous media access (UMA) allows seamless access to multimedia content on any type of devices, especially mobile terminals and hand-held devices. Such user platforms often add unique constraints on media consumption – limited user time, low transmission bandwidth, low power, low-resolution display, etc.

Such unique constraints call for innovative solutions in content analysis, different from techniques developed for other applications such as multimedia search and retrieval. In a search and retrieval context, emphasis is often placed on detection/annotations of semantic objects or events. Techniques culled from pattern recognition, statistical modeling, and information retrieval have been proven fruitful in several video domains, such as news and sports.

In the UMA context, the focus is different. One of the main objectives is to minimize the required resource consumption so that compact, relevant content can be selected, delivered, and presented with the highest quality in response to user's specific interest.

In this paper, we discuss two important technical areas related to content analysis for UMA applications – *summarization* and *adaptation*. The objective of summarization is to maximize the information rate from system to user in media access activities. Summaries of media content allow users rapidly grasp important findings and/or highlights of long video programs before selecting specific piece for in-depth follow-up manipulations.

Once specific items of content are determined, media adaptation aims at transformation of selected content in order to meet various resource constraints such as format, bandwidth, user time, etc. Figure 1 shows the architecture including summarization and adaptation in the data flow in UMA applications.

In the remaining sections, we elaborate and identify important research issues in video summarization and adaptation in the context of UMA.

2. VIDEO SUMMARIZATION

Summarization exists at multiple levels.

- The *semantic-level summaries* typically include major findings of events, objects, and scenes with high-level interpretation according to domain knowledge. For example, a summary of a baseball game may include the final outcome, scoring events/players, and highlights of the game. Summaries of a news video may include brief description of important events or major stories reported in a news program.
- The *syntactic-level summaries* typically include efficient organization of syntactic structures of a video program. For example, a syntactic summary for news video may include a hierarchical organization with multiple levels corresponding to story, episode, and shot.

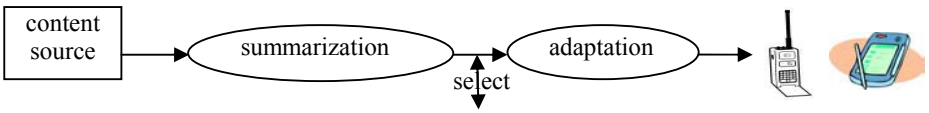

Figure 1. A *summarization-select-adaptation* content processing architecture for UMA applications.

A summary for a baseball program may include multiple levels corresponding to pitches, plays, players, and innings.

The combination of semantic and syntactic summaries provides powerful functionalities for UMA applications. The semantic summary allows users to quickly grasp the major findings and events, while the syntactic summary provides efficient structural indexes, to which each semantic finding can be linked. For example, a sports highlight event mentioned in the semantic summary may be linked to a specific inning or play included in the syntactic summary.

Open Issues:

In recent years, there have been noticeable progresses in syntactic-level video indexing, such as shot segmentation, scene segmentation, sports play detection, and news story segmentation.

Two main challenging issues remain.

- Given any arbitrary video collection, can we discover the underlying syntactic structures automatically without resorting to manual definition and domain knowledge? The capability of unsupervised automatic discovery is important for achieving scalability so that extensive manual labors are not needed when extending solutions to new video domains. In [1], we presented unsupervised pattern discovery techniques based on Hierarchical HMM, statistical model adaptation, and automatic feature selection. Some promising results were shown when the techniques were applied to structured domains such as sports. However, many important issues arise in this emerging fruitful research area.

- Recognizing the difficulty in semantic-level annotation of unconstrained video, an emerging research trend focuses on fusion of multi-modal signals and exploration of their rich spatio-temporal characteristics. Such approaches have shown interesting promises in detecting generic multimedia objects (e.g., monologue), events (e.g., conversations, sports highlight), and locations (e.g., outdoor). However, there is still a significant gap between such generic detectable concepts and the high-level semantic information needed in producing semantic summaries. In view of this, a practical remedy is to incorporate knowledge about the content, user profile, and application task in order to infer high-level semantics from detectable generic concepts. For example, in sports, fusing of visual object recognition, audio/speech recognition, and text recognition provides opportunity for inferring high-level semantics such as the game status, events, and highlights. A key question arises – how do we develop a systematic methodology to incorporate such knowledge for general applications, instead of ad hoc customized solutions.

3. VIDEO ADAPTATION

In UMA environments, media content may have to be adapted into a new one with different format, resolution, or duration, in order to meet the resource constraints or preferences of users. In developing the adaptation methods, there are multiple dimensions involved – adaptation *operations*, *required resources*, and *media quality* (namely *utility*). In each dimension, a myriad of options exist. For example, media adaptation operations may be in the forms of reducing spatial resolution, spatial quality, temporal frame rate, or sequence duration; resources may be defined based on network bandwidth, display size, power consumption, or user time; and utility may be measured in terms of signal distortion, perceptual quality, or high-level comprehensibility. In [2], we presented a unified conceptual framework for modeling relationships between these dimensions, and for finding the optimal adaptation solutions given specific constraints of resource or utility.

One key issue in finding the optimal adaptation is how to estimate the required resources and resulting content utility of any given adaptation operation. The current prevalent approach is to adopt approximate analytical models of the operation and, together with statistical models of the signals, to estimate the required resources and resulting signal quality. Despite its effectiveness in some source coding problems, such an analytical approach is not applicable to general adaptation problems due to the complexity and nonlinearity of the adaptation operations involved.

Content analysis provides an alternative direction with great promise. It's reasonable to assume that any given adaptation operation has similar effects of resource/utility on contents of similar characteristics. For example, video scenes with similar scene complexity, motion patterns, and object activities may be assumed to behave similarly (in terms of utility-resource) under the same adaptation operation. In [3], our experiments with MPEG-4 encoded video confirmed such strong correlation between content characteristics and resource-utility behaviors. We showed that optimal tradeoffs of frame rate dropping and spatial quality reduction can be accurately predicted using computable video features (e.g., motion, spatio-frequency features) and statistical classification techniques. We conjecture such content-based prediction methods can be generalized to other adaptation operations and

video coding formats. Current issues that remain critical are how to define adequate metrics for utility measurement and how to make the predictions tolerant to variations in implementing the same video adaptation operation.

4. CONCLUSION

Content analysis plays a critical role in developing two major technologies for UMA applications – summarization and adaptation. In summarization, the goal of content analysis is to achieve fully automatic discovery of syntactic structures, link the discovered structures to semantics, and to automatically generate concise, meaningful semantic-level summaries. In adaptation, content analysis provides a promising, new direction for solving a core problem in finding optimal adaptation method under heterogeneous constraints. In both areas, promising results have been shown and ample opportunities exist for further technological advancement.

5. REFERENCES

1. L. Xie, S.-F. Chang, A. Divakaran and H. Sun, "Unsupervised Mining of Statistical Temporal Structures in Video," Book Chapter in *Video Mining,* A. Rosenfeld, *et al*, Eds, Kluwer Academic Publishers, June 2003 .
2. S.-F. Chang, Optimal Video Adaptation and Skimming Using a Utility-Based Framework, Tyrrhenian International Workshop on Digital Communications (IWDC-2002), Capri Island, Italy, Sept. 2002.
3. Y. Wang, J.-G. Kim, and S.-F. Chang, Content-based utility function prediction for real-time MPEG-4 transcoding, ICIP 2003, September 14-17, 2003, Barcelona, Spain.