Content-based Video Indexing for Sports Applications using Integrated Multi-Modal Approach

Dian Tjondronegoro
School of Information Technology
Deakin University &
School of Information Systems
Queensland University of Technology
Australia
dian@qut.edu.au

Yi-Ping Phoebe Chen
School of Information Technology
Deakin University
Melbourne, Australia
phoebe@deakin.edu.au

Binh Pham
Faculty of Information Technology
Queensland University of Technology
Brisbane, Australia
b.pham@qut.edu.au

ABSTRACT
To sustain an ongoing rapid growth of video information, there is an emerging demand for a sophisticated content-based video indexing system. However, current video indexing solutions are still immature and lack of any standard. This doctoral consists of a research work based on an integrated *multi-modal* approach for sports video indexing and retrieval. By combining specific features extractable from multiple audio-visual modalities, generic structure and specific events can be detected and classified. During browsing and retrieval, users will benefit from the integration of high-level semantic and some descriptive *mid-level features* such as whistle and close-up view of player(s).

Categories and Subject Descriptors
H.3.1. [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Abstracting Methods*.

General Terms
Algorithms, Experimentation

Keywords
Sports video indexing, multi-modal event detection

1. INTRODUCTION
Triggered by technology innovations, there has been a huge increase in the utilization of video, as one of the most preferred types of media due to its content richness, for many significant applications. To sustain an ongoing rapid growth of video information, there is an emerging demand for a sophisticated content-based video indexing system. However, current video indexing solutions are still immature and lack of any standard. One solution, namely annotation-based indexing, allows video retrieval using textual annotations. However, the major limitations are the restrictions of pre-defined keywords that can be used and the expensive manual work on annotating video. Another solution called feature-based indexing allows video search by low-level features comparison such as query by a sample image. Even though this approach can use automatically extracted features, users would not be able to retrieve video intuitively, based on high-level concepts. This predicament is caused by the so-called semantic gap which highlights the fact that users recall video contents in a high-level abstraction while video is generally stored as an arbitrary sequence of audio-visual tracks.

To bridge the semantic gap, this doctoral will demonstrate the use of domain-specific approach which aims to utilize domain knowledge in facilitating the extraction of high-level concepts directly from the audiovisual features. The main idea behind domain-specific approach is the use of domain knowledge to guide the integration of features from multi-modal tracks. For example, to extract goal segments from soccer and basketball video, slow motion replay scenes (visual) and excitement (audio) should be detected as they are played during most goal segments. Domain-specific indexing also exploits specific browsing and querying methods which are driven by specific users/applications' requirements. Sports video is selected as the primary domain due to its content richness and popularity. Moreover, broadcasted sports videos generally span for hours with many redundant activities and the key segments could make up only 30% to 60% of the entire data depending on the progress of the match.

This doctoral consists of a research work based on an integrated *multi-modal* approach for sports video indexing and retrieval. By combining specific features extractable from multiple (audio-visual) modalities, generic structure and specific events can be detected and classified. During browsing and retrieval, users will benefit from the integration of high-level semantic and some descriptive *mid-level features* such as whistle and close-up view of player(s). The main objective is to contribute to the three major components of sports video indexing systems. The first component is a set of powerful techniques to extract audio-visual features and semantic contents automatically. The main purposes are to reduce manual annotations and to summarize the lengthy contents into a compact, meaningful and more enjoyable presentation. The second component is an expressive and flexible indexing technique that supports gradual index construction. Indexing scheme is essential to determine the methods by which users can access a video database. The third and last component is a query language that can generate dynamic video summaries for smart browsing and support user-oriented retrievals.

2. SYSTEM ARCHITECTURE
The significance of the work lies equally in the approach as a guiding framework and the delivered tools as a validation of the

MM'05, November 6–11, 2005, Singapore.
ACM 1-59593-044-2/05/0011.

approach. Figure 1 depicts the proposed system architecture. The following sets of results are to be presented:

- A number of techniques to extract some mid-level features which can be used to detect generic highlight events from various sports genres. The features set will be used as the basis of other tools
- A Statistical approach for detecting and classifying specific events such as soccer and basketball goals. This approach utilizes a more definitive scope of detection and universal set of feature.
- An integrated summarization scheme that constructs a more complete and self-consumable summaries that can be browsed effectively
- An indexing scheme that supports gradual updates on the data model which combines semi-schema and object-relationship modeling concepts. A query language has been explored to test the benefits of this indexing by constructing some user-oriented retrievals and dynamic summaries construction.

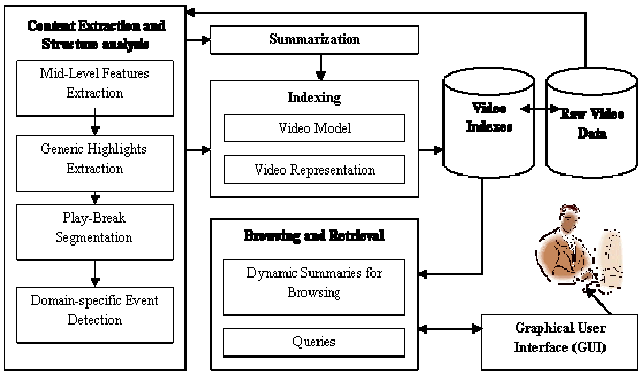

Figure 1. System Architecture

3. MID-LEVEL FEATURES EXTRACTION

Low-level features in video channels such as shape, color, texture (visual), volume, pitch (aural), and keywords (text) are generally not sufficient to support intuitive video retrievals. Despite the fact that the extraction process can be fully automated, there is a semantic gap between users and these features. In order to bridge this gap, a typical pattern of the occurrence of specific features in a particular sports video is normally used to detect domain-specific events. For example, heuristic rules can be used to integrate *crowd cheer*, *score display* and *change in motion direction* for detecting basketball goals [2]. However, sports video can have different styles of presentation. For example, sounds from an excited crowd are usually found during interesting events in soccer but they only exist after each play is stopped in a tennis game. At the same time, they also depend on the broadcasters. For an instance, one broadcaster may use a small text display on the corner of all frames to keep viewers up-to-date with the score-line while another broadcaster may only use medium-sized texts to show the current score-line every time a goal is scored. The main challenge is to design algorithms that can overcome the amount of variation in low-level features by developing a set of mid-level features.

4. DETECTION OF SPORTS EVENTS

Highlights are generically the interesting events that may capture user attentions. While unclassified highlights are good for casual video skimming, domain-specific highlights will support more useful browsing and query applications. For example, users may prefer to watch only the goals. It has become a well-known theory that the high-level semantics in sport video can be detected based on the occurrences of specific audio and visual features which can be extracted automatically. Another alternative is object-motion based which offers a higher level of analysis but requires expensive computations. For example, the definition of a goal in soccer is when the ball passes the goal line inside of the goal-mouth. While object-based features such as ball- and players-tracking are capable of detecting these semantics, specific features like slow-motion replay, excitement, and text display should be able to detect the goal event more efficiently or at least help in narrowing down the scope of the analysis.

To date, there are two main approaches to fuse audio-visual features. One alternative, called *machine-learning* approach, uses probabilistic models to automatically capture the unique patterns of audio visual feature-measurements in specific (highlight) events. For example, Hidden Markov Model (HMM) can be trained to capture the transitions of 'still, standing, walking, throwing, jumping-down and running-down' states during athletic sports' events, which are detected based on color, texture and global-motion measurements [3]. Another alternative for audio-visual fusion is to use manual heuristic rules. For example, Ekin et al [1] detect goals by examining the features that exist within video-frames between the global shot that causes the goal and the global shot that shows the restart of the game. Our main aim us to take advantage of some statistical phenomena of audio-visual features in different highlights to design the rules for highlight classification, thereby reducing subjective domain-knowledge and manual observation.

5. REFERENCES

[1] Ekin, A. and Tekalp, M. Automatic Soccer Video Analysis and Summarization. *IEEE Transaction on Image Processing*, *12* (7). 796-807.

[2] Nepal, S., Srinivasan, U. and Reynolds, G., Automatic detection of 'Goal' segments in basketball videos. in *ACM International Conference on Multimedia*, (Ottawa; Canada, 2001), ACM, 261-269.

[3] Wu, C., Ma, Y.-F., Zhang, H.-J. and Zhong, Y.-Z., Events recognition by semantic inference for sports video. in *Multimedia and Expo, 2002. Proceedings. 2002 IEEE International Conference on*, (2002), 805-808.