Content-Based Scalable Sports Video Retrieval System

Huang-Chia Shih and Chung-Lin Huang
Institute of Electrical Engineering,
National Tsing Hua University,
HsinChu, Taiwan, R.O. C.
E-mail: clhuang@ee.nthu.edu.tw

ABSTRACT

This paper presents a content-oriented video retrieval system which is capable of handling high volumes of content as well as various functionality requirements. It allows the audience to access the video contents based on their different interests of selected video program. The retrieval system consists of a content-based scalable access platform for supporting content-based scalable multi-functional video retrieval (MFVR). The retrieval methodology is based on different content semantic layers. In the experiments, we demonstrate the effectiveness of our method in generating scalable access of video content.

I. INTRODUCTION

The research interest of the digital video browsing has increased enormously over the last decade, due to the rapid growth of video streaming on the World-Wide Web. To access the useful information and reduce the transmission cost, the development of video indexing and retrieval mechnism has become a popular research topic [1, 2]. However, the current mechnisms do not provide sufficient analysis of the video content, so that it does not provide user the useful information. The users can only use the try-and-error searching based on the video indexing. They cannot have an overview of the entire video, neither can they access the summary of the designated video. Revelence feedback [3] mechanism is needed to correct the retrieval mechanism.

Many multimedia retrieval systems have been proposed [4-6], however, most of them provide the key selections (*e.g.*, key frame and key shot) as an index for users to select and browse. Most of them apply the MPEG-7 architecture [7, 8], such as a set of independent or related description. The MPEG-7 has tried to standardize the media access methods, which provide the automatic techniques to access the video based on its content applied for video indexing and retrieval, classification, summarization, and understanding. MPEG-7 has DS (description scheme) to describe the relationships between different variations of video sources to provide adaptive selection under delivery, different platform and user preference conditions. This paper introduces a content scalable access platform for content-based scalable multi-functional video retrieval (MFVR).

II. CONTENT-BASED HIERARCHICAL MODEL

The video content is divided into different layers based on their different semantic significance. As shown in Figure 1, we divide the video semantic contents into four layers:

- **The video clip layer** of which the information header is called the *video clip frame*. It consists of the video clip ID (ID_v) and the raw video data.
- **The object class layer** of which the information header is the *object frame*. It describes the players of interest (POIs) in the video. In each object frame, there is an object ID (ID_o) and pointers which link the video clips to the corresponding POI. Two different object pointers may point to the same video clip.
- **The action class layer** of which the information header is the *action frame*. It describes the action of the POI. The action frame consists of the action ID (ID_a) and linkage to at least one object ID (ID_o). The linkage relates the action among multiple objects.
- **The conclusion layer** of which the information header is the *conclusion frame*. It describes the summarization of the video sequence. The conclusion frame consists of certain semantic summarization of a set of actions. Each conclusion frame consists of semantic ID (ID_c) and a set of ID_a .

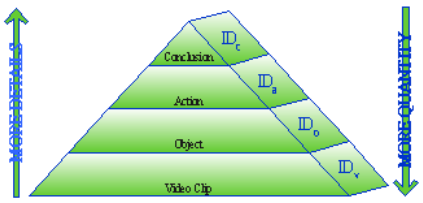


Figure. 1 Hierarchical model of video content

The content-based hierarchical model describes the video by the four-layer frames. Each frame in each layer (except the *video clip frame*) has pointers to the related

frames in the lower and the upper layer. Each frame in the upper three layers has a pointer to the related *video clip frame* in the bottom layer. The pointers is formulated by the joint appearance probability, and the *ID* in each frame is determined by the occurrence function Ω. Assume an object *A* appears in certain video clip, it is denoted by $\Omega(ID_{o=A})=1$. The MFVR scheme allows the client to retrieve different detailed content video of which the retrieval paths are shown in Figure 2.

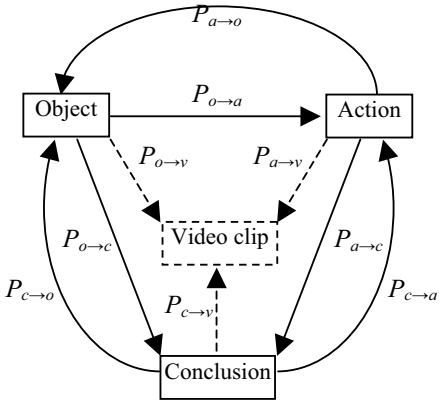


Figure. 2 Paths of MFR for four class elements

As shown in Figure 2, the pointer $P_{start\to end}$ indicates the relation between the *start* and the *end* frames. There are three content-oriented MFVR schemes:

1) ***Object-oriented.*** The client may watch his player of interest (POI) in the video by selecting the designated object frame. Starting from the object frame, he may retrieve all the related frames in other layers.
2) ***Action-oriented.*** The client may watch certain action of play by selecting the action frame which will link to the related object frames involved in this action.
3) ***Conclusion-oriented:*** The client may also retrieve the overall results of the games such as scoring, strike-out, and home-run in baseball game video.

III. CONTEN-BASED SCALABLE MULTIFUNCTION VIDEO RETRIEVAL SYSTEM

The MFVR system consists of many versatile video retrieval schemes. In the following, we will illustrate the MFVR system and its functional blocks.

A. System overview

MFVR is a content-scalable video retrieval system which is designed for different purposes. First, to choose the desired video program from the categorized video archive list, MFVR will provide the key frames and the ID of different semantic layer for the client to choose. Before retrieving the video sequence, the system performs two steps: the visual feature analysis and the content analysis (see Fig. 3).

The key frame detection will extract the significant frames in the video shot. Shot-boundary detection exploits the correlation between the frames in the same shot by calculating consecutive frame dissimilarities and finding the most difference value that indicates the start of a new shot. MFVR is based on two basic modules: the visual feature analysis and the content analysis. The former consists of (1) the object extraction which will identify the foreground object extraction and field feature; (2) the semantic feature modeling for semantic understanding [9] and (3) notice box (NB) region identification. The latter consists of another three analyzers: role analyzer, action analyzer and conclusion analyzer to generate the content ID (*i.e., ID_o, ID_a* and ID_c).

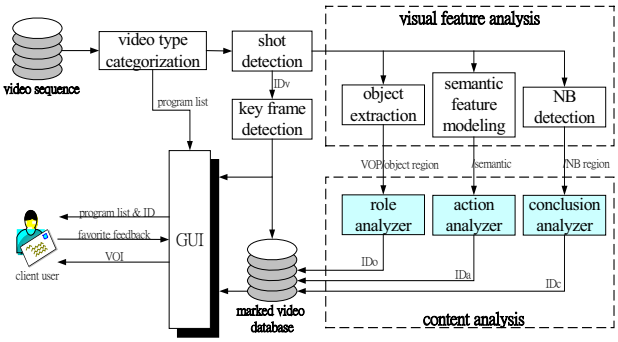


Figure. 3 Overview of content-scalable MFVR system

B. POI Identification

For certain sports video, we find some activity regularity of the POI. Most of the time, the POI moves within certain area of the field during the game, and does certain specific actions. From the location and action regularity of the POI, we may identify his role in the game. These regularities have to be consistent with domain knowledge for MVFR to do the role identification. The domain knowledge denotes the field location of the current video shot based on the analysis of the field lines and the field color.

With the player list which specifies the role of each player and his identification in database, once we identify the role of the POI, we may know his identification. We apply the field domain knowledge to determine whether the extracted object is a POI or not. If we know the role of certain athlete, then we look up the mapping table to obtain his personal data, history data, batting order, and role in the games. Given the specific video sport type, mapping table (including the personal data, batting order, and roles of the players) and field domain knowledge, the POI identification algorithm is described as follows:

Step 1: Video analysis for object segmentation and labeling.

Step 2: Field line/circle/corner detection.

Step 3: Comparison the results of Step 2 with the known field layout of certain sport program to generate the field domain knowledge.

Step 4: From field database to find the extracted object is a POI that fits the field domain knowledge.

Step 5: Assign the role information to the labeled POI based on the template matching for specific field position in the game.

Step 6: Find the ID and the personal data of the identified POI from mapping table.

C. Action Categorization

For a sports video, the MSN [9] system may be used to extract the low-level evidences and then translate the input video into high-level semantic meaning. MSN is a modified BBN structure which can be applied for the action categorization and highlight detection in the baseball video. Based on the Bayes' rule, the posterior probability can be expressed by the joint probability, which can be further expressed by conditional probability and prior probability. Using the *chain rule*, a joint probability distribution can be reduced to a set of conditional probability and a prior probability.

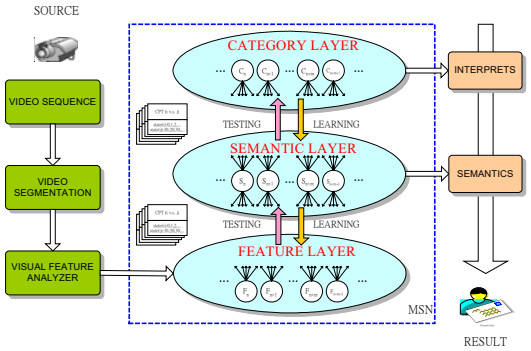


Figure. 4. Action detection based on MSN.

We apply the MSN to generate the action class shown in Figure 4. Top-down learning process will assign the conditional probability table (CPTs) for the linkage between each node pair, whereas the bottom-up testing process will inference the video action based on the low-level features and semantic features.

D. Conclusion Determination

In a regular sports video, to present an instantaneous fact of the game, the program director, normally, will use a so-called notice-box (NB) to display the critical outcomes in the screen. In MFVR, the conclusion frame represents certain outcomes for the video sequence. Since the same action may lead to the different outcomes, we need to identify the NB. Fixed style NB usually exhibits the more ordinary outcomes such as strike-ball count, out number, inning information in a baseball game. Normally, the fixed style NB always appears in a certain region of video frame. Therefore, we may apply the pixel-based color similarity to measure the absolute difference between two frames to extract the static NB region. The similarity between the i^{th} and the $(i - j)^{th}$ frame for pixel n, $S_n(i, j)$ is given as,

$$S_n(i,j) = \frac{1}{N}\sum_{m=1}^{N} \omega_m \left| C_i^m[n] - C_{i-j}^m[n] \right|$$

where ω_m are the weights for the distance measure of color component, N denotes the number of color components, and C_i^m is the value of the i^{th} frame for the same color component. We compare $S_n(i, j)$ with a certain threshold T_s pixel-by-pixel. If $S_n(i, j) < T_s$, then the current pixel is labeled as the candidate pixel in NB region.

IV. EXPERIMENTAL RESULTS

Here, we will demonstrate a prototype of the content-based scalable MFVR system. The image frame of the testing sequence is full color image is 352×240, and its frame rate is 30 f/s. We have selected more than 10 hours (630 min.) from five soccer TV programs of England Premier League (EPL), two soccer TV programs of UEFA Cup, and about 1100 video shots from four baseball competitions, MLB, CPBL, JPBL, baseball game of world cup in 2002 (WDCUP), and the baseball game of Olympic 2004. We use UPMOST mpeg-1 encode card to grab the video clips without causing too much noise and blurring effects on the image recording from TV programs.

A. Fixed style Notice-Box detection

Because the transmutation of fixed NB is almost static, we consider tracking the moving pixel for color similarity under predefined threshold T_s, On the other hand, we can observe that there is significant reduction when the scene change occurs. Consequently, we evaluate the pixel-based color similarity to find the static region (fixed NB).

Table 1. Results of fixed NB detection

Games	Preceding frame	Ts =60
MLB		
		
WDCUP		
EPL		

Table 1 shows the results in the case of $T_s=60$ for RGB color space with $\omega_R=\omega_G=\omega_B=0.9$ demonstrating that our algorithm is compatible for a lots types of sport videos.

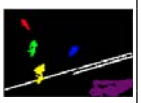
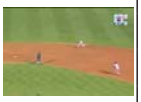
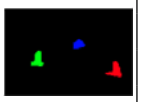
B. Role identification

The object extraction is applied for extracting the POI, the referee, and other objects which are labeled with different colors as shown in the column 2 of Table 2. The field line markers and the grass color are not treated as object but are used to determine the field location. The field domain knowledge and shot type are used to find the possible roles of the POIs in different field and the event shot, as shown in column 3 of Table 2. Using template-matching mechanism, we may identify the role of the POI and his personal data.

C. Experimental Results

Here, we illustrate the implementation of MFVR system. In Figure 5, we represent a user interface of MFVR. Suppose the user selects a player "*Chen (53)*" and he does not care about what is the action and conclusion for "*Chen*" in 1st program list (*i.e.*, World Cup 2001). Then, the system responds with some video at the right-side panel (*i.e.*, the candidate shots) as shown in Figure 5. User can also use the selections at the left-bottom of the panel list to access the video based on the content semantics. The sports video may be query by different semantics as (1) who are players, (2) where in the field they play, (3) what action do they mean, and (4) when it occurs.

Table 2. Results of role identification

Original frame	Object labeling	Field Knowledge Shot type --- Candidate role	Role
		<Striking zone> <Pitching shot> --- <Pitcher>, <Batter> <Catcher>, <Main referee>, <Others>	Batter Pitcher Catcher Main referee
		<1st base> <Defense shot> --- <Runner>, <1st baseman> <Pitcher>, <Batter>, <1st referee>, <Others>	Runner Pitcher 1st Baseman Batter
		<2nd base> <Defense shot> --- <Runner>, <2nd baseman> <Batter>, <2nd referee>, <Others>	Runner 2nd Referee Shortstop 2nd Baseman

V. CONCLUSIONS

We have proposed a content-based video retrieval system. It not only contains the content-scalable scheme that is adaptive for different subjective client user but also supports different video scalable content retrieval. We also illustrate the methodology of content-scalable retrieval based on different content semantics.

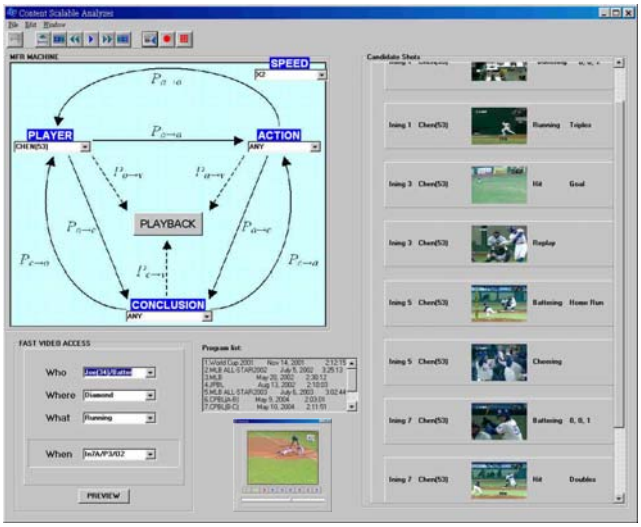


Figure 5. The user interface of MFVR for baseball game.

REFERENCE

[1] D. Zhong and S. F. Chang, "Spatio-temporal video search using the object-based video representation," in *Proc.* IEEE-ICIP, Santa Barbara, Oct. 1997.

[2] M. R. Naphade, I. Kozintsev, and T. S. Huang, "A Factor Graph frame work for demantic indexing," IEEE Trans. on CAS for VT, pp.40-52, Jan. 2002.

[3] Y. Rui, Thomas S. Huang, Michael ortega, and Sharad Mehrotra, "Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval," IEEE Trans. on CSVT, Vol. 8, No. 5 Sep. 1998.

[4] A. Doulamis and N. Doulamis, "Optimal Multi-Content Video Decomposition for Efficient Video Transmission over Low-Bandwidth Networks," in *Proc.* IEEE ICIP, 2002.

[5] Andrew Graves and Mounia Lalmas, "Video Retrieval using an MPEG-7 Based Inference Network," ACM SIGIR'02, Aug. 11-15, 2002.

[6] Tatiana Almeida Souza Coelho, Pavel Pereira Calado, Lamarque Vieira Souza, Berthier Ribeiro-Neto, and Richard Muntz, " IEEE Trans. on Knowledge and Data Engineering, Vol. 16, No. 4, Apr 2004.

[7] ISO/IEC, JTC1/SC29/WG11, "MPEG-7: Overview (version 8)," N4980, July 2002.

[8] S. -F. Chang, T. Sikora, and A. Puri, "Overview of the MPEG-7 Standard," IEEE Trans. on CAS for Video Technology, special issue on MPEG-7, June 2001.

[9] H. C. Shih and C. L. Huang, "Detection of the Highlights in Baseball Video Program," in *Proc.* ICME 2004.