Content-Scalable Analysis for Video Indexing and Retrieval

Huang-Chia Shih

Department of Electrical Engineering, Yuan Ze University,
135 Yuandong Rd., Taoyuan, Taiwan, R.O.C.
hcshih@saturn.yzu.edu.tw

Abstract. Video content-scalability for video indexing and retrieval is proposed. Recently, the demand for content-based multimedia applications is increasing even beyond the capabilities of best-effort transmission networks. Therefore, the trend is toward constructing a content-oriented multimedia server that is capable of handling high volumes of content as well as of fulfilling high performance and various functionality requirements. For instance, in sports programs, audience can request any video clip including any player of the game. This paper introduces a content-scalable access platform for supporting video scalability and role of object identification. First, we describe a hierarchical content analysis method. Second, the framework of content-scalable video indexing with different content semantic class is illustrated.

Keywords: Content scalability, video indexing strategy, role identification, content-based video scalability, hierarchical access.

1 Introduction

The research interest of digital videos browsing has increased enormously over the last decade, due to the rapid growth of video streaming on the World-Wide Web. Many scalable video coding technologies utilizing the spatial and temporal features have been developed [1], [2]. Video scalability can be classified three classes based on the type of scalable unit including spatial, SNR and temporal. But the content always is neglected.

To access the useful information and reduce the transmission cost, the development of video indexing and retrieval mechanism has become a popular research topic [3], [4]. However, the current mechanisms do not provide sufficient analysis of the video content, so that it does not provide the correct information for the user. The user can only fetch what he need from the pre-categorized content through video indexing technology. The user cannot have the overview of the entire video, neither can he realize the representative of the fetched video. Therefore, the iterative retrieval mechanism has been proposed for the user to interactively grade the correctness score of the retrieved picture and send back to the retrieval system. Then the system can adjust its retrieval mechanism to fit what the user really need.

In [5], the video information has been divided temporally into four levels: frame level, key-frame level, shot level, key-shot level，they use the genetic algorithm (GA) to identify the key frame and the key shot. User can manually select the target shot from a set of video shots in the key-shot level, and then he selects the key frames in video shots. In other words, they creating a multi-resolution map from the coarse to the fine resolution, but their description "content-based" still based on the low-level features which to decompose the video performed by minimizing a cross correlation criterion. Although, the user can rapidly to seek the interesting video clip or frame, the shortcoming is not only on any contribution for apprehending the semantic meaning of content because they just to extract the key frame/shot according to the low-level features, but also lacks of the flexibility due to the pre-segmented and pre-classified processes cannot be ensure to comfort for all user request.

In this paper, we proposed a content-scalable video indexing and retrieval system expect to handle the user retrieving based on their real preference.

2 Hierarchical Content Analysis

Due to the mass multimedia data, the available internet bandwidth, and the limited interest of the client, the intelligent video server needs to select and deliver the video of interesting (VOI) from the video program to the client. However, first, the client needs to select his VOIs from video summary. Video summarization research has focus on temporal video segmentation and analysis to select the key elements, such as key shot, key frame, and key object. The advantage of using the key elements to represent the entire video is reducing the viewing time of video for the client. Unfortunately, there is no universal algorithm that can produce the global key elements that fit all clients. There are no globally satisfied video summarization results. We need to develop an effective video retrieve scheme that can be used to deliver the VOIs to the client based on the video content rather than the internet bandwidth only.

2.1. Regularities of Pointing

The hierarchical strategy describes the video by applying the hierarchically structured four-class frames. As shown in Fig. 1, each frame in each class has a pointer to the related frames in the in-coming or out-going class. Each frame in the semantic three classes has a pointer to the related video clip frame in the video clip class. The pointers are 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 $\Omega(ID_{o=A})=1$. For instance, the object frame in the object class has the information of the player *ID* and a set of pointers to the related event frames in event class representing its presence in the video program (as shown in Fig. 1, the pointer P_{oe} indicates the relation between the object frame and the event frame). The event frames have different fidelity to represent the significance of the class in the

video. Since there are other players in the event frame, the fidelity represents its role in the video. The other pointer P_{oc} connects the relation between the object frame and the context frame. The context frame records the outcome of the player's action (i.e., event) in the video which is either a strike-out or a scoring for baseball game.

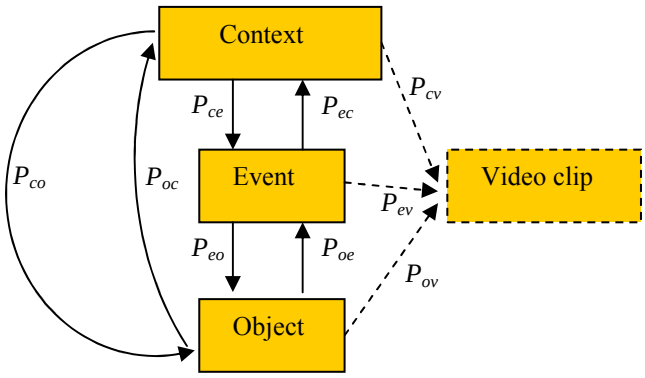

Fig. 1. Pointing schematics of retrieving in-between the four class elements

There are three scenarios for retrieving the tagged video data shown in the follows:

- 1) ***Object-oriented:*** The client may like to watch the portion of the sport video of one of his favorite players in the game, he may select the designated object frame from the object class. Start from the object frame, he may select all the video clips of the player based on the connection pointers, e.g., P_{ov}. If he is interested in the performance of that player in the sport video program, he may specify the event of the player, and then the pointer of the object frame may links to the designated event frame, of which the pointer may relate to the video clip by the linkage, such as $P_{oe} \rightarrow P_{ev}$. For instance, the client may like to watch the scoring video clip of the specific player. By using the set of linkage $P_{oe} \rightarrow P_{ec} \rightarrow P_{cv}$, the client may access the related video clips.
- 2) ***Event-oriented:*** The client may be interested in certain action video clip, he may select the event frame, of which the pointers will link to the related object frames involved in this event. First, he may access all of object frames (of all players) related to the event frame (*i.e.,* P_{eo}). Second, he may retrieve the context frame after the event frame (*i.e.*, P_{ec}). Finally, he may retrieve the video clips related to the specific event (*i.e.*, P_{ev}).
- 3) ***Context-oriented:*** In the last scenario, the client may retrieve the overall results of the games such as scoring, strike-out, and home-run. He may select the context frame directly which will link backward to the video clip frames (*i.e.,* P_{cv}), more detailed event frames (*i.e.*, P_{ce}), or the specific player involved in the contextual outcome (*i.e.*, P_{co}).

2.1. Shot-level Highlight Detection

Prior to determining the key-frame, the key shot will be quickly selected using the information from the scoreboard template. There are two different kinds of key shot: scoreboard-*appeared key shot* and c*ontent-changed key shot*. The former is defined as the shot taken whenever the scoreboard appears, whereas the latter is taken when the scoreboard content changes.

In this paper we assume that the scoreboard template does not change during the entire video program so that we may use the scoreboard color model to identify its presence, even though the scoreboard may be transparent. The similarity measure between model h_R and the potential scoreboard is formulated by the *Mahalanobis distance* as $d(h_i^{Mscb}, h_R)=(h_i^{Mscb}-h_R)^{\mathrm{T}}\mathbf{C}_{\mathrm{m}}^{-1}(h_i^{Mscb}-h_R)$, where h_i^{Mscb} denotes the color distribution of the potential scoreboard region in the ith frame. If a shot is being detected that may change the game status, or if the scoreboard appears, then that shot will be selected as the key shot.

3 Content-Scalable Video Indexing Framework

3.1. System Overview

Content-scalable is most compatible for different scope's client user, when they want to retrieve the database based on their personal subjectivities. Firstly, to choose the desired video program from categorized video archive list, the key frame and the *ID* of different semantic class will be presented to the client. Next, client selects the favorite retrieval conditions in multiple functionality items then feedback the interesting item to indexing system. Before retrieving the video sequence, the system performs two steps: the visual feature analysis and the content analysis. The video program classification and shot boundary detection also be needed. In addition, the key frame detection is developed to extract the most significant frame that can help the client user to decide his prefer video. Shot-boundary detection exploit the correlation between the frames belonging to the same shot by calculating consecutive frame dissimilarities in certainty features (e.q., colour, luminance histogram) and finding for major difference values that should hint the start of a new shots. In visual feature analysis, we evaluate some essential features from low-level visual domain including the video object plane (VOP) or foreground region extraction, inference model needed for semantic understanding [6] and notice-box region identification for conclusion analysis. For content analysis, we have developed three analyzers: role analyzer, event analyzer and context analyzer, to get the content information (i.e., ID_o, ID_a and ID_c). In the paper, we focus on how to extract the role information for every VOP.

3.2. Role Identification

How to identify the player is one of the main subjects of the video retrieval system. In general, people are more interested in enjoying watching their favorite player in the game. Therefore, how to identify the specific player in the entire video is the critical portion of the entire system. Unfortunately, the system can not accurately identify the specific player but it can extract certain visual features (*e.g.,* shape, color, texture etc.) from the video sequence. In MPEG-4 video sequences, we have differentiated distinct objects of frame. Each object is described by the VOP. Here, we assume we have the object information. The next step is how to identify the role of the object in the video sequence. Here we may identify the object in the sports video sequence based on the following approach.

In regular sports video, once a player enters the court, the program director normally will use a so-called notice-box (NB) to introduce the new player. Based on the NB, we may identify the object in the video. First, we find the location of the NB. There are two difference kinds of appearance of the NB: (1) fixed style, (2)floating style (as shown in Fig. 2). Fixed style NBs are fixed on certain region of the screen. They show the current score, the strike-ball count, the current inning, or time information. The frequency of their appearance is high and the significance of the information is lower than the floating style NB. However, it provides the information of the conclusion class, which will be discussed in the next section. On the other hand, the floating style NB appears when a certain event occurs. It appears at the bottom of the screen. To provide the information of the substitute player, the foul player, and the penalty called by the referee. This NB does not appear as often as the previous NB. They appear in the screen as the gradual scene change. In this paper, we using the edge information to obtain the rough position of NB, and then use the OCR technology to identify the text recognition inside the NB. This procedure is well known "Video OCR" mechanism [7], [8]. Then the system may to recognize the player's identification and his scoring record in this game. We adopt the algorithm [9] for text region type classification and recognition.

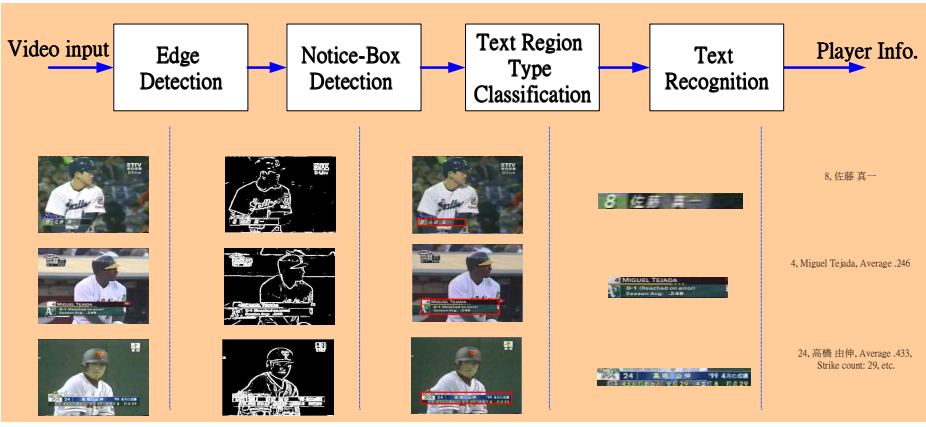

Fig. 2. The example of role identification from Notice-Box.

4 Conclusions

We have proposed a content-scalable video indexing and retrieval system for supporting video scalability and role of object identification. 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.

5 ACKNOWLEDGMENT

This research has been supported by National Science Council (NSC) of Taiwan, under grant number NSC-100-2628-E-155-007- MY2.

References

1. Katata H., Ito N., and Kusao H.: Temporal-Scalable Coding Based on Image Content. IEEE Trans. Circuit and System Video Technology, 7(1), 52--59, (1997).
2. Domanski M., Luczak A., and Mackowiak S.: Spatio-Temporal Scalability for MPEG Video Coding. IEEE Trans. Circuit and System Video Technology, 10(7), 1088--1093 (2000).
3. Zhong D. and Chang S. F.: Spatio-temporal video search using the object-based video representation. in Proc. IEEE-ICIP, Santa Barbara, Oct. 1997.
4. Naphade M. R., Kozintsev I., and Huang T. S.: A Factor Graph frame work for semantic video indexing," IEEE Trans. on CAS for VT, pp.40-52, Jan. 2002.
5. Doulamis A. and Doulamis N.: Optimal Multi-Content Video Decomposition for Efficient Video Transmission over Low-Bandwidth Networks. IEEE Conference on Image Processing, 2002.
6. Shih H.C. and Huang C.L.: Detection of the Highlights in Baseball Video Program. In Proc. IEEE-ICME, 2004.
7. Zhong Y., Zhang H.J., Jain, A.K.: Automatic caption localization in compressed video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(4), 385--392 (2000).
8. Sato, T., Kanade, T., Hughes, E.K., and Smith, M.A.: Video OCR for digital news archive. IEEE International Workshop on Content-Based Access of Image and Video Database, 3 Jan. 1998.
9. Zhang D. and Chang S.F.: Event detection in baseball video using superimposed caption recognition. In Proc. ACM Multimedia, 2002.