VIDEO SUMMARIZATION FOR LARGE SPORTS VIDEO ARCHIVES

Yoshimasa Takahashi, Naoko Nitta, and Noboru Babaguchi

Graduate School of Engineering, Osaka University
2-1 Yamadaoka, Suita, Osaka 565-0871 Japan
{takahashi,naoko,babaguchi}@nanase.comm.eng.osaka-u.ac.jp

ABSTRACT

Video summarization is defined as creating a shorter video clip or a video poster which includes only the important ascenes in the original video streams. In this paper, we propose two methods of generating a summary of arbitrary length for large sports video archives. One is to create a concise video clip by temporally compressing the amount of the video data. The other is to provide a video poster by spatially presenting the image keyframes which together represent the whole video content. Our methods deal with the metadata which has semantic descriptions of video content. Summaries are created according to the significance of each video segment which is normalized in order to handle large sports video archives. We experimentally verified the effectiveness of our methods by comparing the results with man-made video summaries.

1. INTRODUCTION

In recent years, the development of technology has diversified services for the multimedia content, especially the video media. In particular, the technology to quickly search and browse only the information we want from large-scale video archives is crucial. Video summarization is one of the most promising technologies. The objective of video summarization is to create a shorter video clip or a video poster that maintains as much semantic content of the original video streams. Since semantic content of videos is difficult to automatically extract, as an alternative way, we focus on the metadata, which describes the content of videos, to automatically generate a video summary.

Let us describe related work of video summarization. Smith et al.[1] proposed a method of generating a video skim. They extracted significant information from video such as audio keywords, specific objects, camera motions and scene breaks by integrating text, audio, and image analysis. Hanjalic[2] proposed a method for extracting highlights from a sport TV broadcast by detecting strong excitements evoked in a TV viewer by the content of a video. Peker et al.[3] presented a video skimming method where the playback speed was varied based on the visual complexity for an effective fast playback. As a different form of summaries, Uchihashi et al.[4] presented a method of making video posters in which the image size can be changed according to an importance measure. Chiu et al.[5] presented Stained-Glass visualization. The idea of Stained-Glass visualization is to find regions of interest in the video and to condense their keyframes into a tightly packed layout by filling the spaces between the packed regions. However, these methods only use low-level features and do not consider the semantic content, and also the time length of the summary and the number of keyframes to be displayed can not be changed freely. Moreover, since the generated summary is not semantically structured, users still have to view the whole video to search a specific scene.

In this paper, we propose a content-based video summarization method for large sports video archives using metadata which is given to the video media beforehand. A certain number of video segments are selected according to the significance of play scenes. Here, the significance of play scenes is normalized in order to handle several videos at the same time for large sports video archives. The quality of a video summary depends on whether the information which a user wants is included in the summary. Therefore, we consider creating a video summary which fits the length specified by a user and includes as many important video segments as possible. We also try to generate the summary in the form of a video poster which arranges semantically structured image keyframes in a 2-dimensional plane.

2. METADATA FOR SPORTS VIDEOS

Metadata is the data to describe the content, quality, condition, and other characteristics of the data and includes semantic information. MPEG-7[6] has been recently standardized to describe the metadata for videos. In this paper, we assume the metadata, which is described with MPEG-7, is given to videos beforehand.

Sports games generally have tree structures according to their genres, and a sports video can be structured based on the structure of the corresponding sports game. For exam-

0-7803-9332-5/05/$20.00 ©2005 IEEE

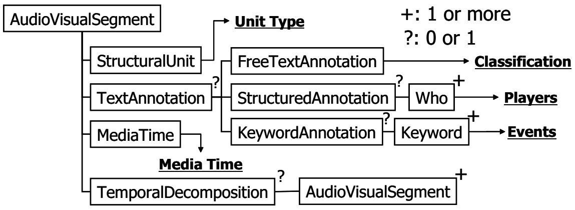


Fig. 1. Composition of metadata

ple, a whole baseball game is composed of several innings, an inning is composed of several at-bats, an at-bat is composed of several plays, and a play is composed of several shots. Note that a play corresponds to a pitcher throw for baseball. These tree structures are described in the metadata for sports videos. Additionally, for each play scene, five items of information, 1) the unit type, 2) the classification, 3) the players, 4) the events, and 5) the media time are described as shown in Fig.1.

3. VIDEO SUMMARIZATION

One of the strong demands for videos is to understand their content in a short time. Video summarization is one of the solutions. Video summarization is defined as creating a shorter video clip or a video poster which includes only the important scenes in the original video streams. Therefore, each play scene should be ranked according to its semantic importance[7]. In this section, we propose a method of making a summary by ranking play scenes using metadata.

3.1. Play Scene Selection

The highlights of a game should be generated based on the significance of play scenes. Furthermore, play scenes are selected so that the time length of the video summary does not exceed the time specified by a user.

3.1.1. Significance of Play Scenes

Each play scene is given a score based on three components: the play ranks, the play occurrence time, and the number of replays[8].

1) Play Ranks

The rank-based significance degree of a play scene p_i, s_r $(0 \leq s_r \leq 1)$, is defined as

$$s_r(p_i) = 1 - \alpha \cdot \frac{r_i - 1}{5} \tag{1}$$

where r_i denotes the rank of the *i*th play scene p_i and α $(0 \leq \alpha \leq 1)$ is the coefficient to consider how much the difference of the rank affects the significance of play scenes.

2) Play Occurrence Time

We define the occurrence-time-based significance degree of a play scene p_i, s_t $(0 \leq s_t \leq 1)$, as

$$s_t(p_i) = 1 - \beta \cdot \frac{N - i}{N - 1} \tag{2}$$

where N is the number of all play scenes and β $(0 \leq \beta \leq 1)$ is the coefficient to consider how much the occurrence time affects the significance of play scenes.

3) Number of Replays

We define the number-of-replays-based significance degree of a play scene p_i, s_n $(0 \leq s_n \leq 1)$, as

$$s_n(p_i) = 1 - \gamma \cdot \frac{n_{\max} - n_i}{n_{\max}} \tag{3}$$

where n_i denotes the number of replays of the *i*th play scene p_i, $n_{\max}$ is the maximum number of n_i, and γ $(0 \leq \gamma \leq 1)$ is the coefficient to consider how much the number of replays affects the significance of play scenes. As a consequence, significance degree of a play scene p_i is given by

$$s(p_i) = s_r(p_i) \cdot s_t(p_i) \cdot s_n(p_i) \tag{4}$$

Changing the parameters of α, β, and γ enables us to control the composition of the video summary. Larger α can emphasize the significance of the play ranks. The other parameters behave in a similar manner.

In this paper, more than one videos are handled at the same time. Since it is difficult to compare the significance of play scenes in different videos, z-score is used in order to conform the distribution of the significance degree in all videos. Z-score $z(p_i)$ is calculated as

$$z(p_i) = \frac{s(p_i) - \overline{s(p_i)}}{SD} \tag{5}$$

where $\overline{s(p_i)}$ denotes the average of $s(p_i)$ and SD is the standard deviation.

3.1.2. Selection of Highlights

Here, how to generate a video summary based on the significance of play scenes is described. For a video clip, when the time length of a video summary L is given to the system with a function $\varphi\big(l(p_i)\big)$ $\big(0 < \varphi\big(l(p_i)\big) \leq l(p_i)\big)$ which changes the length of a play scene p_i, the problem can be formulized as follows.

select subset $P' = \{p_j \mid j = 1, 2, \ldots, k\}$ $(1 \leq k \leq N)$
from play scene set $P = \{p_1, p_2, \ldots, p_N\}$,
subject to $\displaystyle\sum_{p_j \in P'} z(p_j) \longrightarrow \max$

$$\sum_{p_j \in P'} \varphi\big(l(p_j)\big) \leq L$$

where N denotes the total number of play scenes, $z(p_i)$ denotes the z-score of the significance of a play scene p_i, and

$l(p_i)$ denotes the time length of a play scene p_i. Thus, we can define this problem as the combinational optimization problem with constrained conditions. For a video poster, L denotes the area of the space to display image keyframes, l denotes the area of each keyframe, and φ denotes the function which changes the area of each keyframe.

Basic Method: The basic method selects play scenes in the order of their significance. First, we sort out the play scene set P in the order of significance. Next, we select play scenes in sequential order from the first play scene in P until the sum of the length of the selected play scenes exceeds the time specified by the user.

3.2. Visualization

The form of a video summary is classified into two types: a video clip and a video poster. In a video clip, the time length of the original videos is temporally compressed. In a video poster, image keyframes are presented in a 2-dimensional plane. We describe the details about the video clip and the video poster as follows.

3.2.1. Video Clip

The characteristic of this method is that the total length of the summary can be flexibly changed according to the time specified by a user. We propose two more methods to select play scenes based on their significance.

Greedy Method: The greedy method first arranges play scenes in the order of significance per unit time. Next, the play scenes are selected in the arranged order until the sum of the length of the selected play scenes exceeds the time specified by a user.

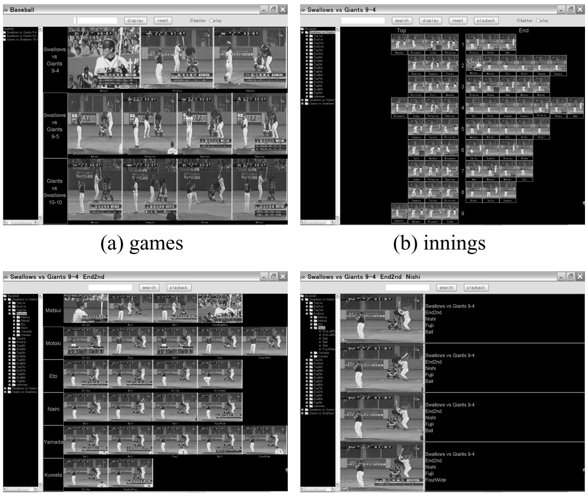
Play-Cut Method: According to the time specified by a user, the length of play scenes is cut short dynamically. Therefore, we call this method "Play-Cut Method." The play scene set P is arranged in the order of significance. Next, we select play scenes in sequential order from the first play scene in P until the sum of the length of the selected play scenes exceeds the time specified by the user. Here, the following calculation is performed in order to put bounds to the length of a play scene.

$$\varphi\big(l(p_i)\big) = \min\,[\,l(p_i),\; l_{th} + \delta \cdot L'\,] \qquad (6)$$

where l_{th} denotes the threshold of the minimum time required for the user to grasp the content of a play scene, and L' denotes the current remaining time after subtracting the total time of the selected play scenes from L, and δ is the coefficient to consider how much L' affects the length of play scenes.

3.2.2. Video Poster

We also propose a spatial visualization system which provides image keyframes each of which represents a play scene in the summary. Fig.2 shows the interface. Sports games generally have tree structures according to their genres, and a sports video can be structured based on the tree structure of the corresponding sports game. In the video poster, the tree structures of the sports games are displayed in the left side as game trees and the keyframes of the scenes corresponding to the selected tree node are displayed in the right side. The first frame of each scene is displayed as the keyframe.



(a) games (b) innings

(c) at-bats (d) plays

Fig. 2. Interface

First of all, keyframes of important scenes for each game are displayed in a row as shown in Fig.2(a). Additionally, only the keyframes which agree with the keywords such as players or events can be displayed. All innings in a game are displayed by selecting the game, all at-bats in an inning is displayed by selecting the inning, and all plays in an at-bat is displayed by selecting the at-bat. In Fig.2(b), each row represents an inning. The innings line up from top to bottom and the keyframes for each inning line up from left to right in the temporal order. In Fig.2(c), each row represents an at-bat. The at-bats line up from top to bottom and the plays for each at-bat line up from left to right in the temporal order. In Fig.2(d), the plays line up from top to bottom in the temporal order. Users can efficiently access the scene they want to see by hierarchically tracing the game tree. Other functions of our system are as follows.

Display of Highlights: With the video poster, users can directly specify the number of keyframes to be displayed as shown in Fig.3. Moreover, only the keyframes suitable for keywords can also be displayed. In addition, only the keyframes with high significance degrees can be displayed as a highlight.

Playback of each Play Scene: Users can view only specific play scenes by clicking their keyframes.

Annotations of Video Content: Since it is difficult to understand the semantic content only by looking at the keyframes,

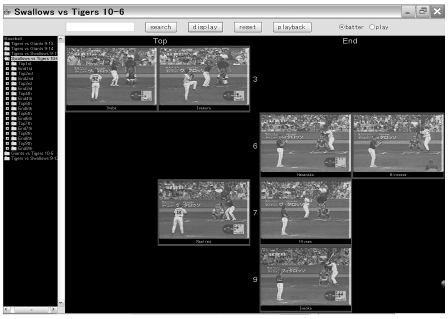


Fig. 3. The display of highlights

the system also displays annotations about the players and the events.

4. EXPERIMENTS

We prepared 5 baseball videos broadcasted by three different TV stations with an average length of 3 hours and 30 minutes. The summaries generated by our methods were compared with the summaries (of 5 games) which were broadcasted as highlights by the same TV stations as the original videos. We assumed that the play scenes in the summaries broadcasted on TV are the correct answer set for video summarization. We evaluated the results with the recall (=(the number of the play scenes included in both TV and our summary / the number of the play scenes included in the TV summary) × 100) and the precision (=(the number of the play scenes included in both TV and our summary / the number of the play scenes included in our summary) × 100).

The value of the parameters were experimentally determined as $\alpha = 0.8$, $\beta = 0.1$, $\gamma = 0.3$, $l_{th} = 14$, and $\delta = 0.02$. The comparative results between the summaries generated with our methods and the summaries broadcasted on TV are shown in Table 1. We first conducted the basic, greedy, and play-cut method setting the length of the summary to 120 seconds, then conducted the play-cut method setting the length of the summary to the same length (55-110 seconds) of each TV summary because the recall of the play-cut method was the highest of all methods.

The details of the results of three methods are as follows: The summaries generated by the basic method included only a few play scenes. This result is due to the limitation of the space defined by the length of the summaries specified by users. Although the play scenes in the summaries were mostly the important play scenes, each of them included many redundant shots taking up space for other important play scenes. On the other hand, although the summaries generated by the greedy method had sufficient number of play scenes, they also included many unimportant play scenes. The play-cut method was able to get the best summaries. This method obtained only the important play scenes by getting rid of the redundant shots which were included in the summaries generated by the basic method, leaving space for other important play scenes.

The current system serves only for baseball videos; however, the framework of the system itself is applicable to other types of sports videos with similar game structures. For that purpose, the value of the parameters should be changed depending on the types of sports videos. Consequently, some systematic way to determine the value of the parameters should be contrived.

Table 1. Comparative Results

method	# of plays in both	# of plays on TV	# of plays in our summary	recall	precision
1.basic method	2.8	7.8	3.8	44%	80%
2.greedy method	2.0	7.8	8.2	29%	26%
3.play-cut method 1	5.0	7.8	8.0	66%	63%
4.play-cut method 2	5.0	7.8	5.8	66%	83%

5. CONCLUSION

In this paper, we proposed an automatic content-based video summarization method using metadata for large sports video archives. We also presented two visualization systems for the video summary: video clip and video poster, formulated problems specific to each type, and proposed a method for generating the video summary of arbitrary length and dealing with several videos at the same time. As a result of experiments with baseball videos, we obtained only the significant play scenes with the recall rate of 66% and the precision rate of 83% compared with the summaries broadcasted on TV. As a future work, personalization in making video summaries should be considered to meet users' preferences. The browsing system should also be improved by extracting keyframes which best describe the semantic content of the scenes.

6. REFERENCES

[1] M.A.Smith and T.Kanade, "Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques," Proc. IEEE CVPR'97, pp.775-781, June, 1997.

[2] A.Hanjalic, "Generic Approach to Highlights Extraction from a Sport Video," Proc. IEEE ICIP 2003, Vol.1, pp.1-4, September, 2003.

[3] K.A.Peker and A.Divakaran, "Adaptive Fast Playback-based Video Skimming using a Compressed-Domain Visual Complexity Measure," Proc. IEEE ICME 2004, June, 2004.

[4] S.Uchihashi, J.Foote, A.Girgensohn, and J.Boreczky, "Video Manga: Generating Semantically Meaningful Video Summaries," Proc. ACM Multimedia'99, pp.383-392, October, 1999.

[5] P.Chiu, A.Girgensohn, and Q.Liu, "Stained-Glass Visualization for Highly Condensed Video Summaries," Proc. IEEE ICME 2004, June, 2004.

[6] J.M.Martinez, "Overview of the MPEG-7 Standard (version 6.0)," ISO/IEC JTC1/SC29/WG11 N4509, December, 2001.

[7] N.Babaguchi, Y.Kawai, T.Ogura, and T.Kitahashi, "Personalized Abstraction of Broadcasted American Football Video by Highlight Selection," IEEE Trans. Multimedia, Vol.6, No.4, pp.575-586, August, 2004.

[8] Y.Takahashi, N.Nitta, and N.Babaguchi, "Automatic Video Summarization of Sports Videos Using Metadata," Proc. IEEE PCM 2004, Vol.2, pp.272-280, December, 2004.