2009 11th IEEE International Symposium on Multimedia

TENNIS VIDEO WITH SEMANTIC SCALABILITY

Jui-Hsin Lai [1] *and Shao-Yi Chien* [2]

Media IC and System Lab
Graduate Institute of Electronics Engineering and Department of Electrical Engineering
National Taiwan University
BL-421, 1, Sec. 4, Roosevelt Rd., Taipei 106, Taiwan
[1] larry@media.ee.ntu.edu.tw, [2] sychien@cc.ee.ntu.edu.tw

ABSTRACT

Scalable video is the research topic to provide different size of video bitstream under different transmission bandwidth. In this paper, the semantic scalability is proposed that provides the scalable videos in semantic domain, and the tennis videos are used as the experiments. Contrary to decreasing the video quality to reduce the bitrates, the lower bitstream size is achieved by abandoning the video contents with less semantic importance. The experimental results show that the proposed semantic scalability provides four levels of the scalable videos and maintains the visual quality in watching the game video. The study of the scalability in semantic domain provides a new aspect for the scalable video.

Index Terms— Content adaptive, scalable video, video rendering

1. INTRODUCTION

Watching sport videos is a popular entertainment in our daily lives. With the enlargement of video resolution and the improvement of video quality, the bitstream size of videos is dramatically increased. As the progressing in broadcasting technologies, people are able to watch the game videos by various devices from televisions, computers to mobile phones. However, people can not enjoy the high-quality videos everywhere due to the limitation of the transmission bandwidth.

To provide the lower size of video bitstream, scalable video coding (SVC), the scalable extension of advance video coding (AVC) [1], is a current standardization project of video compression under different transmission bandwidth. It provided variable bitstream sizes by reducing the video resolution (spatial domain), decreasing the number of video frames (temporal domain), and increasing the quantization parameters (PSNR domain). However, the viewing quality under lower bitstream size was often seriously decreased and can not be accepted by people. For the following works of SVC, Akyol et al. proposed the objective function to choose the best scaling type for each temporal segment that resulted in minimum visual distortion [2]. Nevertheless, the viewing quality under lower bitstream size was still unacceptable. Unlike the approaches of bitstream reduction in SVC, Tang et al. presented a content-adaptive system for streaming the goal events of soccer games over network with low bandwidth limitation [3]. Instead of the low-quality videos, the panoramic field images were used to present the game events under the lower bandwidth. However, the excitement of the games was also decreased due to the presentation by still images. In addition, Wikstrand and Eriksson employed the animations to represent the football game on mobile phones [4]. The presentation with the animations was an interesting concept to reduce the bitstream size, but the concept should be extended to provide more scalable options.

In considering the visual quality, we propose the semantic scalability to provide four levels of the scalable videos. For the implementation of semantic scalability, the proposed video rendering is the key step. By reintegrating these extracted video contents, the video rendering can generate the scalable videos with different bitstream sizes. Contrary to scaling the bitstream sizes in spatial, temporal, and PSNR domains, the scalability provides different video sizes in the semantic domain. The smaller bitstream sizes are achieved by the four appraoches: the reuse of the video background, the abandonment of video clips, the abandonment of video contents, and the video presentation in animations. In other words, the unnecessary video contents are removed in priority of semantic importance, and the visual quality can still be maintained. In addition, the semantic scalability also has the property of adaptive transmission that the bitstream size can be immediately adjusted by changing to other scalable levels. The study of the scalability in semantic domain provides a new aspect for the scalable video.

2. ANALYSIS AND RENDERING OF GAME VIDEOS

The video rendering is the key step to achieve the scalability, and the video contents as the rendered materials are extracted by the video analysis. In this section, we would introduce the video analysis and the video rendering.

978-0-7695-3890-7/09 $26.00 © 2009 IEEE
DOI 10.1109/ISM.2009.13

IEEE computer society

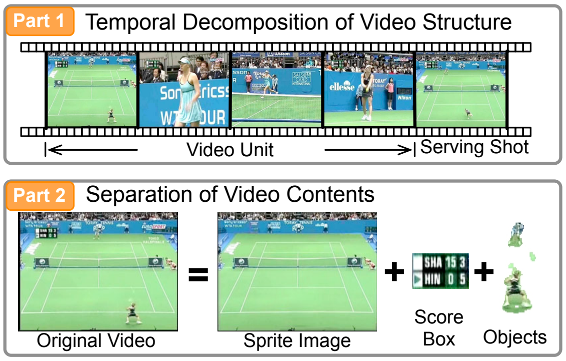

Fig. 1. The video analysis contains two parts: temporal decomposition of video structure and the separation of video contents.

2.1. Video Analysis

The length of a game video is usually several hours. For a long-time video, the video partition can reduce the difficulties in video analysis. In our observation, the tennis videos are repeating the iteration: player serving, game running, and game stop. One serving-running-stop video clip can be seemed as a Video Unit [5]. Each Video Unit, which begins at a serving shot and ends before the next serving shot, usually presents an event in a tennis video. With such a regular structure, the analysis of video structure can be done by finding all serving shots in the game video. Then the whole video is decomposed into Video Units as the Part 1 in Fig. 1.

For each Video Unit, the serving shot contains the complete game information. How to extract the information of serving shots is the important task for video analysis. With the help from the sprite image, the video contents are decomposed into background court and foreground objects like players, ball, ball boys and audiences [6]. Furthermore, the coordinates of foreground objects and the camera translation parameters are also extracted. As the Part 2 in Fig. 1, the video contents of serving shots are fully separated and annotated.

2.2. Video Rendering

The proposed video rendering reintegrates the extracted contents in Section 2.1 and generates the plentiful video according to user's requests [6]. As the illustration in Fig. 2, the sprite image, which is also referred to as the background court, can be processed by inserting the text or advertisements. By modifying the parameters of camera translation, the user can control the motions of the virtual camera with panning, titling, and zooming. Finally, the foreground objects are pasted on the watching view, and the number of the pasted objects can also be controlled by users.

Unlike all video contents in the same layer, the rendered video is composed of several layers like the background court, players, ball, ball bays and audiences. By editing the layered contents, the customized video is rendered and the highlight replays can be automatically generated. For example, with the extracted player trajectories, the virtual camera can focus on the favorite player, and the replay of player tracking is generated. Furthermore, the video rendering also can implement the semantic scalability. For instance, the reduction of video bitstream can be achieved by decreasing the transmission number of the layered contents. The more details of bitstream reduction are described in Section 3.

3. SEMANTIC SCALABILITY

To reduce the bitstream size and maintain the visual quality, we propose four approaches to abandon the video contents in the priority of semantic importance. In other words, the smaller bitstream size is achieved by reducing the unnecessary video contents. The proposed tennis video with semantic scalability has four levels as shown in Fig. 3.

3.1. Video Background Reuse

It can be observed that the tennis court rapidly appears in a game video and has the large ratio of area in the serving shots. If the background court can be reused in the video broadcasting, it will have the obviously decrease in bitstream sizes. So, the bitstream reduction in Level 1 of the Fig. 3 is to individually transmit the background court and foreground objects for the serving shots. Notice that the background court is only transmitted once and reused in the subsequent video. Due to the complexity of the scenes, the non-serving shots are transmitted in the single layer without further video processing. Although the background court is abridged in the subsequently transmission, the rendered video in Level 1 can be identical to the original game video. So, the bitstream reduction in Level 1 is achieved by reducing the redundant transmission of the background court.

3.2. Video Clip Abandonment

To decrease the bitstream size in Level 2, the approach is to abandon the video clips with less semantic importance. From the angle of semantic importance, the serving shots have more important game information than the non-serving shots which are usually the event replays or the zoom-in view on players. Furthermore, the average bitstream sizes of the non-serving shots are greater than the serving shots due to the larger number of scene changes. From the above considerations, the non-serving shots are abandoned and the total bitstream sizes are extremely reduced. For the absence of non-serving shots, the highlight replays in Section 2.2 are rendered to fill the empty time. The highlight replay is the appropriate method to complete the abandonment of the non-serving shots.

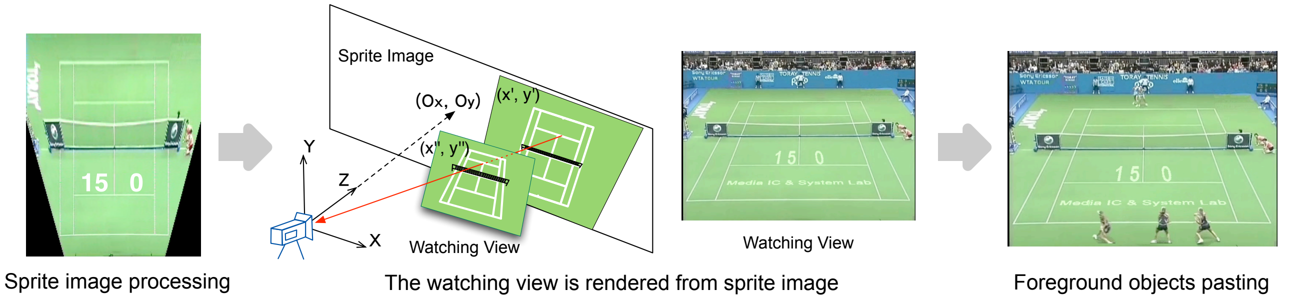

Fig. 2. The proposed video rendering contains three steps: sprite image processing, watching view rendering, and foreground objects pasting.

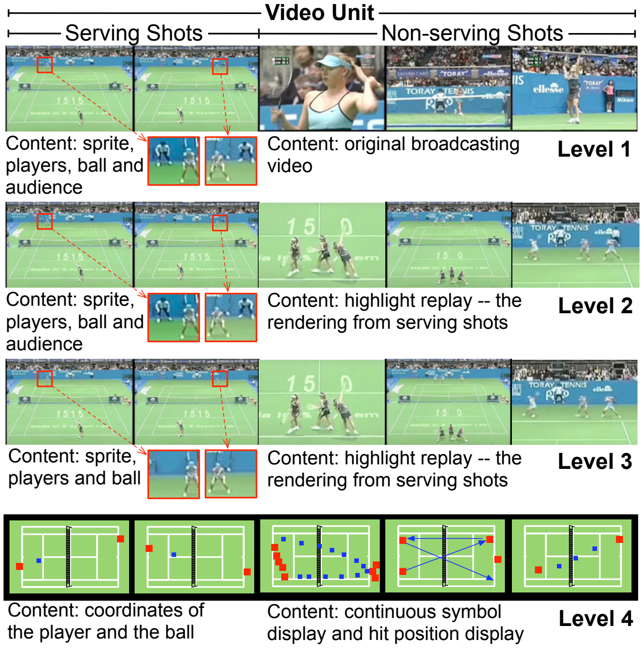
Video Unit
Serving Shots
Non-serving Shots
Content: sprite, players, ball and audience
Content: original broadcasting video
Level 1
Content: sprite, players, ball and audience
Content: highlight replay -- the rendering from serving shots
Level 2
Content: sprite, players and ball
Content: highlight replay -- the rendering from serving shots
Level 3
Content: coordinates of the player and the ball
Content: continuous symbol display and hit position display
Level 4

Fig. 3. The four levels of semantic scalability. The bitstream size in each level is decreased by reducing the unnecessary video contents.

3.3. Video Content Abandonment

To further decrease the bitstream size in Level 3, the approach is to abandon the video contents in the serving shots. In watching the game video, people would pay most attention on the players but pay less attention on the non-attractive foreground objects like the referee, ball boys, and people in the auditorium. However, these non-attractive foreground objects take lots transmission bandwidth. Therefore the non-attractive foreground objects are abandoned to reduce the bitstream size, and only the ball and players are transmitted. To fill the empty time from the absence of the non-serving shots, highlight replays are rendered from the serving shots like the process in Level 2. In addition, there is an interesting effect in Level 3 that all the objects, without the ball and players, are static in the video.

3.4. Video in Animation

To extremely decrease the bitstream size in Level 4, the animation is employed to present the game video. By extending the reduction in Level 3, only coordinates of the ball and players are transmitted. Level 4 is proposed for extremely low transmission bandwidth like the real-time broadcasting on the mobile devices. Although there are no player gestures or other detailed game information, the state of the game is still roughly represented by these coordinates. For the time interval of the non-serving shots, the statistic data, like player trajectories and hit positions, are shown.

4. EXPERIMENTAL RESULTS

The tennis videos with resolution 720x480 are used as the experiments. The proposed semantic scalability can be presented as the four levels in Fig. 3, and the video demo is also available on the website [7]. The average bitrates of each scalable level are shown in Fig. 4.

For the compression of the background court in Level 1, the sprite image is encoded by Lossless JPEG [8], which can achieve 6.453 times of average compression rates. In other words, a sprite image with resolution 1080x720 in RGB formate has 362KB file size. Notice that the sprite image is reused in the video, so the sprite image only needs to be transmitted once in the beginning as the red bin in Fig. 4. For the compression of the foreground objects, all the foreground objects are encoded by the main profile of H.264 video encoder (JM15) [1], and the average bitrates are 261K bits per second(bps). The non-serving shots are also encoded by the main profile of H.264 video encoder (JM15) , and the average bitrates are 1230 Kbps. We can see that the background reuse in the serving shots has the obvious lower bitrates in comparison with the bitrates of non-serving shots.

For the compression bitrates in Level 2, the background court and foreground objects of the serving shots are individually encoded as the same in Level 1. For the abandonment of non-serving shots, the highlight replays are rendered from the serving shots and do not need another transmission data. So, the bitrates in the time interval of the non-serving shots are zero, but the additional cost is the buffer to store the contents of the serving shots. We can see that the abandonment of non-serving shots has the dramatical bitrates reduction in comparison with Level 1. By using the highlight replays to replace the non-serving shots, people have less visual loss in watching the game videos.

For the compression bitrates in Level 3, the background court and foreground objects are individually encoded as the same in Level 2, but the difference is that the foreground objects only include the ball and players. The average bitrates of the foreground objects are only 144 Kbps, which are about the half size in comparison with Level 2. For the abandonment of non-serving shots, the bitrates of the highlight replays are zero as the same in Level 2. We have observed that the viewing quality in Level 3 is almost identical to Level 2 even though the non-attractive foreground objects are abandoned.

For the compression bitrates in Level 4, only the position coordinates of the ball and players are transmitted. Instead of the video frames, positions of the ball and players are displayed on the court map. The data size of these coordinates is only 5.76 Kbps. The statistic data in the time interval of non-serving shots is retrieved from the coordinates in serving shots. The total bitrates in Level 4 are extremely lower than other levels.

The experimental results show that the proposed semantic scalability provides scalable videos with maintaining the visual quality in watching the game videos. Furthermore, the semantic scalability also has the property of adaptive transmission that the bitstream size can be immediately adjusted according to the transmission bandwidth. For example, people can watch the game video in Level 2 and receive some non-serving shots when the bandwidth is available.

5. CONCLUSION

The proposed semantic scalability provides four levels of the scalable videos and maintains the visual quality in watching the game videos. Without decreasing the video quality to reduce the bitrates, the lower bitstream size is achieved by abandoning the video contents with less semantic importance. Although only tennis videos are used to implement the semantic scalability, we will employ more sport videos as the experiments in the future works.

6. REFERENCES

[1] *Recommendation H.264: Advanced video coding for generic audiovisual services*. ITU-T, 2003.

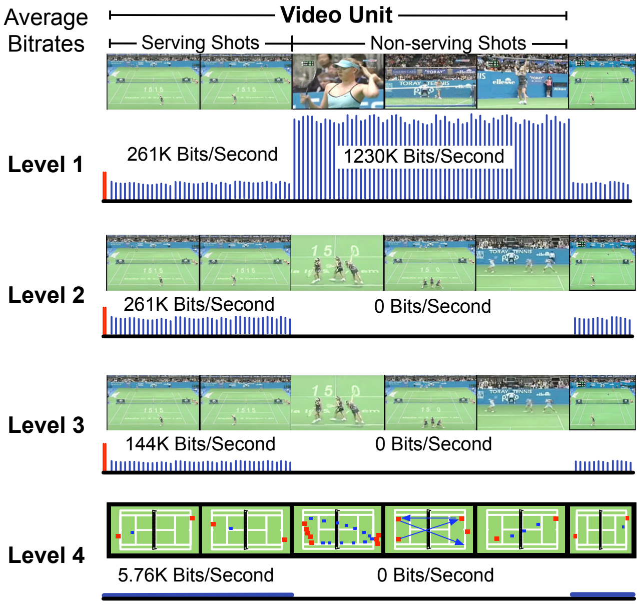

Fig. 4. The average bitrates of each scalable level. The red bin in the beginning is the bitrate of the sprite image, and the blue bins are the average bitrates of video contents.

[2] E. Akyol, A. M. Tekalp, and M. R. Civanlar, "Content-aware scalability-type selection for rate adaptation of scalable video," *EURASIP Journal on Applied Signal Processing*, vol. 1, 2007.

[3] Q. Tang, I. Koprinska, and J. S. Jin, "Content-adaptive transmission of reconstructed soccer goal events over low bandwidth networks," in *Proceedings of the 13th Annual ACM International Conference on Multimedia, MULTIMEDIA '03*, Nov. 2005, pp. 271–274.

[4] G. Wikstrand and S. Eriksson, "Football animations for mobile phones," in *Proceedings of the Second Nordic Conference on Human-Computer Interaction NordiCHI '02*, October 2002, pp. 255–258.

[5] J.-H. Lai and S.-Y. Chien, "Baseball and tennis video annotation with temporal structure decomposition," in *Proceedings of IEEE International Workshop on Multimedia Signal Processing, MMSP 2008*, Oct. 2008, pp. 676–679.

[6] ——, "Tennis video enrichment with content layer separation and real-time rendering in sprite plane," in *Proceedings of IEEE International Workshop on Multimedia Signal Processing, MMSP 2008*, Oct. 2008, pp. 672–675.

[7] [Online]. Available: http://media.ee.ntu.edu.tw/larry/scalable/

[8] *Recommendation JPEG Standard*. ITU-T, 1993.