Author's personal copy

Multimed Tools Appl (2012) 59:585–599
DOI 10.1007/s11042-010-0685-x

Semantic scalability using tennis videos as examples

Jui-Hsin Lai · Shao-Yi Chien

Published online: 29 December 2010
© Springer Science+Business Media, LLC 2010

Abstract With advances in broadcasting technologies, people are now able to watch videos on devices such as televisions, computers, and mobile phones. Scalable video provides video bitstreams of different size under different transmission bandwidths. In this paper, a semantic scalability scheme with four levels is proposed, and tennis videos are used as examples in experiments to test the scheme. Rather than detecting shot categories to determine suitable scaling options for Scalable Video Coding (SVC) as in previous studies, the proposed method analyzes a video, transmits video content according to semantic priority, and reintegrates the extracted contents in the receiver. The purpose of the lower bitstream size in the proposed method is to discard video content of low semantic importance instead of decreasing the video quality to reduce the video bitstream. The experimental results show that visual quality is still maintained in our method despite reducing the bitstream size. Further, in a user study, we show that evaluators identify the visual quality as more acceptable and the video information as clearer than those of SVC. Finally, we suggest that the proposed scalability scheme in the semantic domain, which provides a new dimension for scaling videos, can be extended to various video categories.

Keywords Content adaptive · Scalable video · Video rendering · Video analysis · Scalable video coding

J.-H. Lai (✉) · S.-Y. Chien
Media IC and System Lab, Graduate Institute of Electronics Engineering,
National Taiwan University, BL-421, 1, Sec. 4, Roosevelt Rd., Taipei 106, Taiwan
e-mail: juihsin.lai@gmail.com

S.-Y. Chien
e-mail: sychien@cc.ee.ntu.edu.tw

Author's personal copy

1 Introduction

With improvements in video resolution and quality, the bitstream size of videos has dramatically increased. Given the advances in broadcasting technologies, people are now able to watch videos from devices including televisions, computers, and mobile phones. However, they cannot yet enjoy high-quality videos everywhere because of limitations in the transmission bandwidth.

To provide video bitstreams of lower size, Scalable Video Coding (SVC)—the scalable extension of Advance Video Coding (AVC) [12]—is the current standard for video compression under different transmission bandwidths. SVC provides variable bitstream sizes by reducing the video resolution (spatial domain), decreasing the number of video frames (temporal domain), and increasing the quantization parameters (SNR domain). However, the viewing quality under a lower bitstream size is often seriously compromised and unacceptable to viewers. To improve the viewing quality of videos compressed by SVC, Thang et al. [14] exhaustively reviewed previous studies and pointed out a number of potential solutions to these problems. Among these studies, some have sought an optimized way to discard enhancement layers of different scalability-types to maintain video quality. For example, Wang et al. [15] proposed a general classification-based prediction framework for selecting the most suitable adaptation based on subjective quality evaluations—the first attempt to apply domain-specific knowledge to construct distinct video categories sharing similar scalable operations. Akyol et al. [2] subsequently determined the weights of this objective function for different content types and bitrates using a training procedure with subjective evaluations. Unlike the approaches of bitstream reduction in SVC, Tang et al. [13] presented a content-adaptive system for streaming goal events in soccer videos over a network with low bandwidth limitations. Instead of low-quality videos, panoramic field images were used to present the events under a lower bandwidth. However, the excitement of the games was also reduced because of the presentation of still images. In addition, Wikstrand and Eriksson [16] employed animations to represent football videos on mobile phones. This interesting concept to reduce the bitstream size should be extended to provide more scalable options.

Figure 1a shows the conventional SVC scheme. Previous studies detected video categories and applied different compression schemes to different video categories.

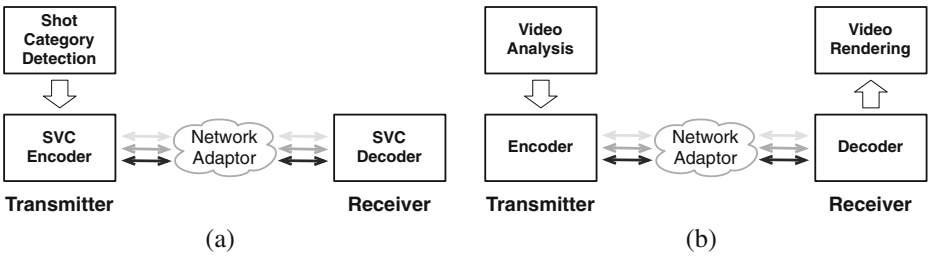

Fig. 1 **a** Conventional scheme of scalable video coding. Previous studies proposed methods for Shot Category Detection to determine suitable scaling options for SVC. **b** The proposed semantic scalability scheme. A video clip is analyzed and labeled as belonging to a particular shot category, after which foreground objects and background scenes are extracted. These extracted video materials are transmitted in order of semantic importance and reintegrated in the receiver

Author's personal copy

For instance, Wang et al. [15] classified videos into three categories based on content complexity and concluded that more complex videos require more bits for spatial details rather than a higher frame rate. Akyol et al. [2] employed soccer videos as examples and classified video shots into four types according to the distance of the shot and the type of motion, after which each shot type was coded with the best scaling option. Note that these previous studies proposed functions to determine suitable scaling options such as spatial, temporal, and SNR scalabilities for different video types and bitrate constraints.

In this paper, we propose a semantic scalability scheme to provide four different levels of scalable videos that preserve visual quality under low video bitrates. To the best of our knowledge, the proposed scalability scheme in the semantic domain is different from schemes proposed in previous studies and highlights a new dimension for scaling videos. Figure 1b shows the processing scheme for semantic scalability. Contrary to scaling the bitstream sizes in the spatial, temporal, and SNR domains, the proposed scalability scheme reduces bitrates by discarding video content in order of priority of semantic importance—a concept which we refer to as semantic scalability.

The process of video annotation and extraction is a key step to implement semantic scalability. First, this process builds background scenes, analyzes shot categories, segments foreground objects, and extracts video information. For some video clips, background scenes occupy a large area of the frame for several seconds. Repeatedly appearing background scenes, which usually comprise a large proportion of a video bitstream, can be recognized as redundant information in video coding. Therefore, the first level of the semantic scalability scheme re-uses background scenes in video coding. Next, different clip categories contain different amounts of semantic information, and different foreground objects also present different degrees of semantic importance. Thus, the second and third levels of semantic scalability discard video clips containing less information in the video transmission and foreground objects with less semantic importance, respectively. To further reduce the transmission bitrate, the fourth level replaces video frames with animations.

The main contributions of this paper are as follows.

- Unlike scaling the bitstream sizes in the spatial, temporal, and SNR domains, the proposed scalability scheme reduces bitrates by discarding video content in order of priority of semantic importance. The results show that despite effectively reducing video bitrates, subjective evaluations reveal that in watching a video, visual quality is better and understanding is clearer than for SVC. Scalability in semantic domain can be seen as a fourth dimension of scalable video coding.
- The nature of semantic scalability is to preserve video content with higher subjective importance under bitrate reduction. Therefore, a key step in the transmission scheme is video analysis, which labels the importance of each clip and object in a video. A corresponding process to reintegrate these clips and objects is required in the receiver. We propose a transmission scheme that includes video analysis, a compression scheme, and video rendering to implement semantic scalability (Fig. 1b).

Several tennis videos are used as examples to demonstrate the semantic scalability scheme. In addition, the concept of semantic scalability is extended to home videos or other sports videos. The paper is organized as follows. Section 2 introduces the

Author's personal copy

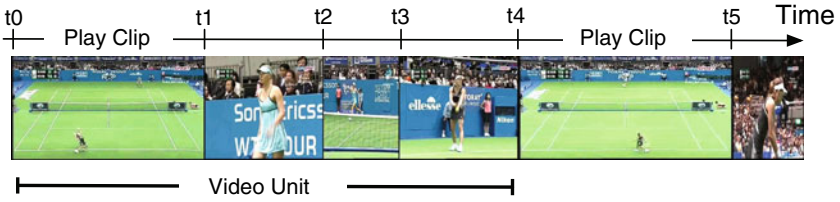

Fig. 2 Video Unit representing a single service-play-stop video clip in a tennis match

methods of video analysis and video rendering. Details of the semantic scalability scheme are presented in Section 3, and experimental results are presented in Section 4. Finally, we offer conclusions and extensions of this work.

2 Video analysis and rendering

2.1 Video analysis

The video of a game (e.g., a tennis match) is usually several hours long. For a long video, partitioning it into shorter clips can reduce the difficulties of video analysis. We have observed that videos of tennis matches (hereafter tennis videos) repeat the iteration: player serving, game running, and game on hold. As shown in Fig. 2, a single service-play-stop video clip can be regarded as a Video Unit. Each Video Unit, which begins with a play clip[1] and ends before the next clip, usually represents an event in tennis videos. Given such regularity, the structure of a video can be analyzed by finding all play clips within it. In previous studies of clip recognition, Lai and Chien proposed the method of template matching by color histograms [5]; Han et al. recognized play clips by calculating the number of white pixels in video frames [3]. Both methods could precisely detect the play clips and decompose the entire video into Video Units.

For each Video Unit, the play clip contains a vast amount of game information such as ball and player trajectories. The primary task of video analysis is to extract this information from the play clips. Segmentation of the ball and players in tennis videos is difficult because the possible camera motions during a match are panning, tilting, and zooming. Lai et al. proposed a method of projecting the video frames onto a sprite plane [7]. This method reconstructed the background scene and effectively segmented the foreground objects, as shown in Fig. 3. To extract players from foreground objects, Han et al. proposed a method using a mean-shift tracking algorithm and blob separation under occlusion of multiple objects [3]. To retrieve the ball from foreground objects, Lai and Chien proposed a method using a Kalman-based motion model to predict and complete ball trajectories [6].

[1] Rallies constitute the play clips in tennis videos, as shown in Fig. 2.

Author's personal copy

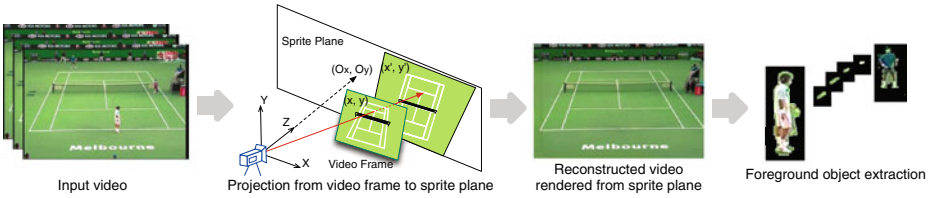

Fig. 3 Process flow for extracting foreground objects. Video frames are projected to the sprite plane, the background scene used to segment foreground objects

2.2 Video rendering

The proposed method of video rendering reintegrates the video content extracted in Section 2.1. As the illustration in Fig. 4 shows, the sprite plane, also referred to as the background scene, can be processed by inserting text or advertisements. By modifying the extrinsic parameters of the camera in (1), a virtual camera can be rendered with panning, titling, and zooming.

$$\begin{bmatrix} x'' \\ y'' \\ w'' \end{bmatrix} = \begin{bmatrix} m_1 & m_2 & m_3 \\ m_4 & m_5 & m_6 \\ m_7 & m_8 & 1 \end{bmatrix} \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix}, \tag{1}$$

where $(x',\ y')$ are the coordinates in the sprite plane, and $(x''/w'',\ y''/w'')$ are the coordinates in the rendering view. Finally, the foreground objects, whose number may depend on the users request, are pasted on the viewing angle.

Unlike videos in the conventional transmission scheme, the rendered video is composed of several layers including the background court, players, ball, ball boys, and moving audiences. By editing the layered contents, a customized video is rendered and highlight replays can be generated by reintegrating these contents. For example, with the extracted player trajectories, the virtual camera can focus on a specific player and generate a replay of the viewing angle focusing on the player. Furthermore, video rendering can also implement semantic scalability. For instance, size reduction of a video bitstream can be achieved by decreasing the number of layered contents transmitted. The details of bitstream reduction are described in Section 3.

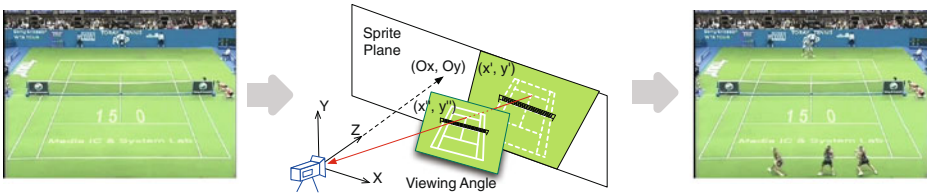

Fig. 4 Three steps of the proposed video rendering: sprite-image processing, viewing-angle rendering, and foreground-object pasting

Author's personal copy

3 Semantic scalability

To reduce the bitstream size and maintain visual quality, we propose four approaches to discard video content in order of priority of semantic importance. In other words, a smaller bitstream size is achieved by reducing the amount of video content transmitted. The proposed semantic scalability scheme has four levels. These are shown in Fig. 5.

3.1 Video background re-use

It can be observed that a view of the tennis court, which covers a large percentage of the area of play clips, abruptly appears in a tennis video. If the background court can be re-used in video transmission, the bitstream size can be reduced. Thus, the bitstream reduction in Level 1 of Fig. 5 is achieved by re-using the background scene. It should be noted that the background court and foreground objects of the play clips are individually transmitted, and the former is transmitted only once and re-used in subsequent videos. Although the background court is abridged in subsequent transmissions, the rendered video in Level 1 is still identical to the original tennis video. Thus, the bitstream reduction in Level 1 is achieved by reducing redundant transmission of the background court. In addition, non-play clips are transmitted in a single layer without further processing because the efficiency of background re-use is insufficient.

3.2 Discarding of video clips

To reduce the bitstream size in Level 2, video clips with less semantic importance are discarded. With regard to semantic importance, play clips have more important

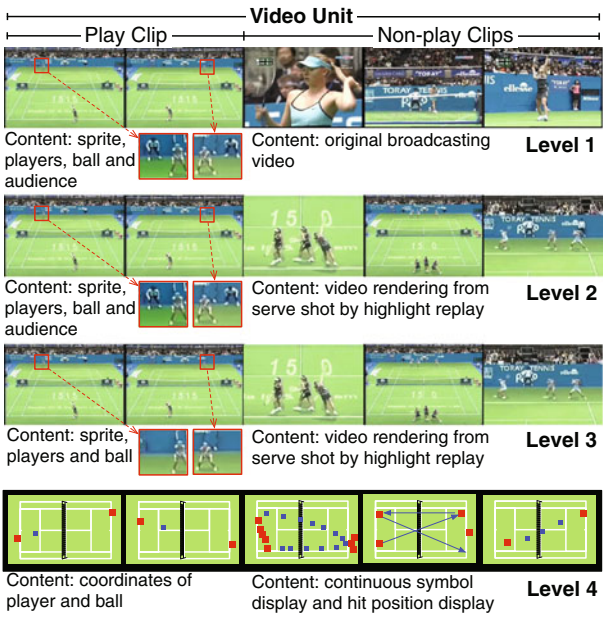

Fig. 5 Four levels of semantic scalability. The bitstream size in each level is reduced by decreasing the amount of video content

Author's personal copy

game information than non-play clips, which are usually event replays or close-up views of players. Furthermore, the average bitstream sizes for non-play clips are greater than those for play clips because of the abrupt changes in scene. Given these considerations, non-play clips are discarded and the total bitstream size is greatly reduced. To fill the absence of non-play clips, highlights from the play clips are played. The highlights can be rendered from play clips that show players running long distances to reach the ball. This is because a situation in which a player forces his/her opponent to run a long distance to return the ball is inherently exciting.

3.3 Discarding of video content

To further reduce the bitstream size in Level 3, certain video content in play clips is discarded. While watching tennis videos, people mostly pay most attention to the players and little to the referee, ball boys, people in the audience, etc. However, the latter non-essential components take up considerable transmission bandwidth. Therefore, they are discarded to reduce the bitstream size, and only the ball and players are transmitted. To fill in the empty non-play time, highlight replays similar to those described in Level 2 can be rendered. An interesting feature of Level 3 is that all objects, except for the ball and players, are static in the video.

3.4 Video in animation

In order to reduce the bitstream size in Level 4, animation is employed to represent the tennis video. The reduction in Level 3 is further extended to transmit only the positions of the ball and players. Level 4 is proposed for extremely low transmission bandwidths. Although lacking in detailed game information such as player postures, the coordinates of the ball and players can be used to roughly represent the state of the game. During the non-play clips in the original video, statistics such as player trajectories and hit positions can be shown. With ball and player trajectories, the rendered computer graphics can provide users with new experiences of watching tennis videos [8].

4 Experimental results

Different tennis videos in Table 1 are used as test videos. Video demonstrations of experimental results are available on a website [10].

Table 1 Several tennis videos used in the experiments

	Game	Video resolution
Video 1	Final of 2007 Australia open men's single	720 × 480
Video 2	Semi final of 2008 US open men's single	720 × 480
Video 3	Semi final of 2009 French open men's single	720 × 480
Video 4	Semi final of 2009 Wimbledon open women's single	720 × 480

Author's personal copy

4.1 Bitstream size at each level

For compression of the background court in Level 1, the sprite plane is encoded by Lossless JPEG [11] to preserve better image quality and achieve compression rates of 6.5 times the average rates. In other words, a sprite image with a resolution of 1080×720 in RGB format has a file size of 362 KB. It should be noted that as the sprite image is re-used in the video, it only needs to be transmitted once at the beginning, as shown by the red bin in Fig. 6. For compression of foreground objects, all are encoded by the main profile of the H.264 video encoder (JM15) [12] with average bitrates of 261 kilobits per second (Kbps). Non-play clips are also encoded by the main profile of the H.264 video encoder (JM15) with average bitrates of 1230 Kbps. We see that background re-use in play clips clearly has lower bitrates than in non-play clips.

For compression of bitrates in Level 2, the background court and foreground objects of play clips are individually encoded as in Level 1. To discard non-play clips, highlight replays that do not require further data transmission are rendered from the play clips. Thus, bitrates in the time interval containing non-play clips in the original tennis video are zero, although the buffer to store the contents of play clips poses an additional cost. We see that by discarding non-play clips, the bitrates are dramatically reduced compared to Level 1. By replacing non-play clips with highlight replays, there is less visual loss in watching the tennis videos.

For compression of bitrates in Level 3, the background court and foreground objects are individually encoded as in Level 2, but with the difference that the latter only include the ball and players. The average bitrates of the foreground objects are only 144 Kbps about the half the size of those in Level 2. By discarding non-play clips, the bitrates of the highlight replays are zero as in Level 2. We have observed that the viewing quality in Level 3 is almost identical to that in Level 2 even though non-attractive foreground objects are discarded.

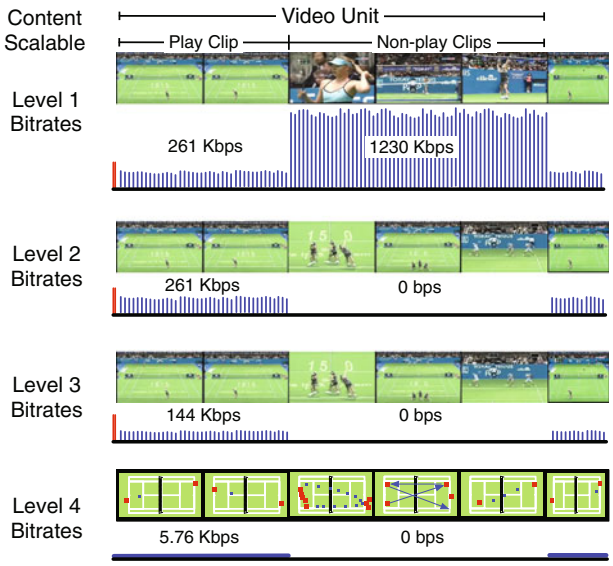

Fig. 6 Average bitrates for each scalable level. The *red bin* at the beginning (*far left*) is the bitrate for the sprite image, and the *blue bins* are the average bitrates for video content

Author's personal copy

For compression of bitrates in Level 4, only the position coordinates of the ball and players are transmitted. Instead of video frames, ball and player positions are displayed on the court map. The data size for these coordinates is only 5.76 Kbps. The statistics presented during the time window emptied of non-play clips and replaced with highlights is retrieved from the coordinates of the play clips. The total bitrates in Level 4 are extremely low relative to those in other levels.

Experimental results show that the proposed semantic scalability scheme reduces the video bitrates by discarding video content in order of priority of semantic importance. Furthermore, the semantic scalability scheme also has the property of adaptive transmission in that the bitstream size can immediately be adjusted according to the transmission bandwidth. For example, viewers can watch the tennis video in Level 2 while receiving some non-play clips when the bandwidth is available.

4.2 Comparison of bitstream size and visual quality

The comparisons of bitstream size and visual quality between the proposed semantic scalability and SVC schemes are shown in Fig. 7. For SVC in the spatial domain, a lower bitstream size is achieved by reducing the video resolution, although visual

Fig. 7 Comparisons of bitstream size and visual quality for SVC in the spatial, temporal, and SNR domains, as well as for the proposed semantic scalability scheme

Author's personal copy

Table 2 Mean opinion score (*MOS*) and standard deviation (*SD*) of compressed videos in the evaluation of visual quality

SVC in spatial domain	SVC in temporal domain	SVC in SNR domain	Proposed scalability in semantic domain
Resolution: 720 × 480 Bitrate: 1020 Kbps MOS: 5 SD: 0	Frame rate: 30 fps Bitrate: 1020 Kbps MOS: 5 SD: 0	QP: 30 Bitrate: 1020 Kbps MOS: 5 SD: 0	Level 1 Bitrate: 895 Kbps MOS: 5 SD: 0
Resolution: 360 × 240 Bitrate: 526 Kbps MOS: 3.32 SD: 0.45	Frame rate: 7.5 fps Bitrate: 525 Kbps MOS: 3.64 SD: 0.78	QP: 40 Bitrate: 327 Kbps MOS: 2.84 SD: 0.50	Level 2 Bitrate: 117 Kbps MOS: 4.76 SD: 0.12
Resolution: 180 × 120 Bitrate: 264 Kbps MOS: 1.98 SD: 0.74	Frame rate: 2 fps Bitrate: 370 Kbps MOS: 2.32 SD: 0.66	QP: 50 Bitrate: 102 Kbps MOS: 1.38 SD: 0.78	Level 3 Bitrate: 64 Kbps MOS: 4.78 SD: 0.22

quality is reduced because of the smaller display regions. For SVC in the temporal domain, a lower bitstream size is achieved by reducing the frame rate; however, visual quality decreases as a result of discontinuous camera motion and player postures. For SVC in the SNR domain, a lower bitstream size is achieved by increasing the quantization parameters (QP); nevertheless, the resulting video is blurred and visual quality is reduced. For the proposed scalability scheme in the semantic domain, a lower bitstream size is achieved by discarding the contents in order of priority of semantic importance without reducing visual quality. Note that the ball boy is eliminated in the second level of scalability in Fig. 7, although viewers may not notice the difference. Four demo videos are available[2] [10], which show detailed visuals of these comparisons.

To evaluate the visual quality of these compressed videos, we also invited twenty subjects, all of whom were graduate students. There are several test methods to evaluate video quality. We employed the Double Stimulus Impairment Scale Method, giving the subject two chances to examine the reference and test sequences prior to providing a response. For example, the subject was required to watch the reference video (with the highest bitrate) for 10 s, rest for 5 s, watch the compressed video for 10 s, and then repeat this procedure. The subjects were instructed to compare the test sequence to the reference sequence and judge the visual quality based on a five point scale, i.e., 5, imperceptible; 4, perceptible but not annoying; 3, slightly annoying; 2, annoying; and 1, very annoying. The mean opinion scores (MOS) and standard deviations (SD) for each of the compressed videos are shown in Table 2, the results of which are also illustrated in Fig. 8. We see that the visual quality of videos compressed by SVC methods was seriously reduced at lower bitrates. In particular, videos with bitrates under 440 Kbps compressed in the spatial and temporal domains made subjects feel slightly annoyed; this annoyance became unacceptable at bitrates under 250 Kbps. In contrast, the visual quality of the proposed method is still preserved and does not decrease as the bitrate declines (Fig. 7). This is because the

[2] http://media.ee.ntu.edu.tw/larry/scalable/

Author's personal copy

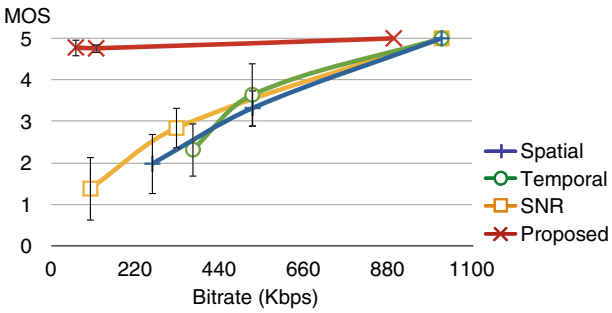

Fig. 8 Mean opinion score (*MOS*) and standard deviation (*SD*) of compressed videos in the evaluation of visual quality

reduced bitrates are due to elimination of information, as in the re-use of background scenes and the discarding of video clips and content.

4.3 Evaluation of video understanding

The same subjects mentioned in Section 4.2 were invited to test their understanding of the game after watching compressed videos. We also employed the Double Stimulus Impairment Scale Method with the same settings as in Section 4.2, and asked the subjects to give scores based on a five point scale, i.e., 5, imperceptible; 4, perceptible but not annoying; 3, slightly annoying; 2, annoying; and 1, very annoying. The MOS and SD for the understanding of each compressed video are shown in Table 3. Note that results of Level 4 in the proposed method are only listed in Table 3 but not in Table 2. This is because the visual quality of the animated video in Level 4 is difficult to compare with that of non-animated videos. Figure 9 illustrates the results in Table 3. We see that it is more difficult to understand videos compressed by SVC methods at lower bitrates. This is particularly true in the case where the ball

Table 3 Mean opinion score (*MOS*) and standard deviation (*SD*) of compressed videos in the evaluation of understanding

SVC in spatial domain	SVC in temporal domain	SVC in SNR domain	Proposed scalability in semantic domain
Resolution: 720 × 480 Bitrate: 1020 Kbps MOS: 5 SD: 0	Frame rate: 30 fps Bitrate: 1020 Kbps MOS: 5 SD: 0	QP: 30 Bitrate: 1020 Kbps MOS: 5 SD: 0	Level 1 Bitrate: 895 Kbps MOS: 5 SD: 0
Resolution: 360 × 240 Bitrate: 526 Kbps MOS: 3.98 SD: 0.55	Frame rate: 7.5 fps Bitrate: 525 Kbps MOS: 4.33 SD: 0.44	QP: 40 Bitrate: 327 Kbps MOS: 2.89 SD: 0.53	Level 2 Bitrate: 117 Kbps MOS: 4.59 SD: 0.35
Resolution: 180 × 120 Bitrate: 264 Kbps MOS: 2.62 SD: 0.82	Frame rate: 2 fps Bitrate: 370 Kbps MOS: 3.76 SD: 0.67	QP: 50 Bitrate: 102 Kbps MOS: 1.97 SD: 0.69	Level 3 Bitrate: 64 Kbps MOS: 4.52 SD: 0.20
			Level 4 Bitrate: 4 Kbps MOS: 2.78 SD: 1.03

Author's personal copy

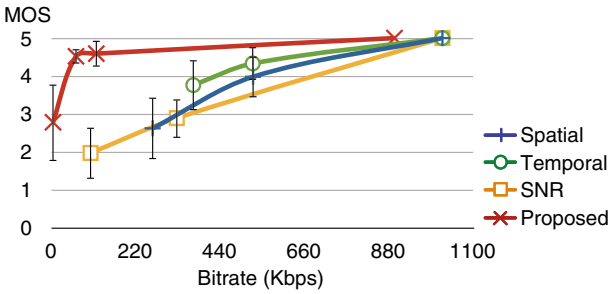

Fig. 9 Mean opinion score (*MOS*) and standard deviation (*SD*) of compressed videos in the evaluation of understanding

disappears at Levels 2 and 3 of the SNR domain because of high QP. In a tennis video, the ball trajectory contains important semantic information, and numerous subjects feel that it is difficult to understand the game without it. Furthermore, in Level 3 of the SNR domain, players were severely blurred and viewers had trouble recognizing a players posture. In contrast, the MOSs for the proposed method only slightly declined as the bitrate decreased because video content with higher semantic importance was still preserved. The results also revealed that the players and ball contain a considerable amount of game information, and that viewers pay most attention to these when watching a tennis video.

4.4 Total user experience

We next designed a subjective test to evaluate the total experience for evaluators: ten with a habit of watching tennis videos and ten others without such a habit. First, the evaluators watched the sequences under bitstream reduction by SVC and then under bitstream reduction by semantic scalability. Subsequently, the evaluators compared the visual quality of the sequences between SVC and semantic scalability under bitstream reduction. Three questions were designed to evaluate visual quality, and evaluators gave a score of 1 to 9 depending on their degree of satisfaction, i.e., 1, very unsatisfied; 3, unsatisfied; 5, no difference; 7, satisfied; and 9, very satisfied. The questions were as follows:

Q.1 Do you think the visual quality of videos is more acceptable and game information is clearer under bitstream reduction by semantic scalability?

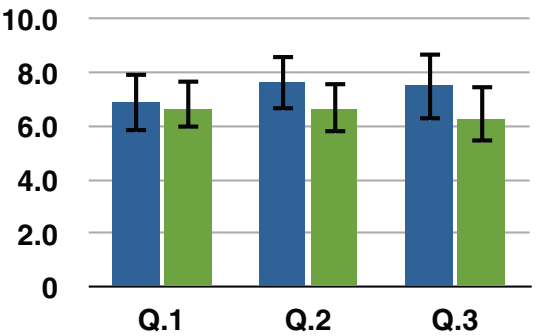

Fig. 10 Results of subjective evaluation. *Blue* and *green bars* are the average scores for evaluators with and without a habit of watching tennis, respectively. The *error bars* show standard deviations

Author's personal copy

Q.2 Do you think semantic scalability is a more practical solution in video broadcasting?
Q.3 Are you willing to watch tennis videos with semantic scalability?

The average scores for the questions, shown in Fig. 10, are all higher than 6 in spite of evaluators with or without the habit of watching tennis. From these results, evaluators identified with the statement that compared to SVC, the visual quality of semantic scalability is more acceptable and game information is clearer. In addition, they were willing to watch tennis videos with semantic scalability functions, and identified with the statement that the proposed semantic scalability scheme is a more practical solution than SVC in video broadcasting. Another interesting phenomenon is that the scores from evaluators with tennis-viewing habits are higher than those from evaluators without such habits. It appears that people who often watch tennis identify closely with the contributions of semantic scalability and prefer to have these functions.

5 Conclusions and extensions

To provide videos with variable bitstream sizes, we propose four levels of scalability in the semantic domain. Without compromising video quality to reduce the bitrates, a lower bitstream size is achieved by discarding video content of low semantic importance. Experimental results show that video bitrates can be effectively reduced by re-using the background scenes, discarding some video clips, reducing the amount of foreground objects, and replacing the video with animation, respectively. Subjective evaluations reveal that the visual quality of videos compressed by our proposed method is better than that of videos compressed by SVC at low bitrates. In addition, understanding is clearer when watching videos compressed by the proposed method than those compressed by SVC at low bitrates. To the best of our knowledge, the proposed scalability scheme in the semantic domain is different from previously studied schemes and provides a new dimension for scalable video coding.

Although we only used tennis videos as examples in this paper, semantic scalability can be extended to further video categories such as home videos and football videos. To extend the concept of a Video Unit to such videos, play clips (in tennis) can be replaced with clips attracting more visual attention [9] and shoot clips, respectively. Video analysis is a key step to achieve semantic scalability; the process of foreground segmentation by projecting video frames to the sprite plane can be applied to various video categories [7]. To extract video information, camera motion can be used as a clue to annotate home videos [1], and previous methods of extracting football and player trajectories are available [4, 17]. With video information and content thus extracted, the semantic scalability scheme can be applied to various video categories.

References

1. Abdollahian G, Taskiran C, Pizlo Z, Delp E (2010) Camera motion based analysis of user generated video. IEEE Trans Multimedia 12(1):28–41
2. Akyol E, Tekalp AM, Civanlar MR (2007) Content-aware scalability-type selection for rate adaptation of scalable video. EURASIP J. Appl Signal Process 7(1):214–225

Author's personal copy

3. Han J, Farin D, de With PHN (2008) Broadcast court-net sports video analysis using fast 3-d camera modeling. IEEE Trans Circuits Syst Video Technol 18(11):1628–1638
4. Inamoto N, Saito H (2004) Free viewpoint video synthesis and presentation of sporting events for mixed reality entertainment. In: Proceedings of the 2004 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology, pp 42–50
5. Lai J-H, Chien, S-Y (2008) Baseball and tennis video annotation with temporal structure decomposition. In: Proceedings of IEEE International Workshop on Multimedia Signal Processing, MMSP 2008, pp 676–679
6. Lai J-H, Chien S-Y (2008) 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, pp 672–675
7. Lai J-H, Kao C-C, Chien S-Y (2009) Super-resolution sprite with foreground removal. In: Proceedings of IEEE International Conference on Multimedia and Expo, pp 1306–1309
8. Liang D, Huang Q, Liu Y, Zhu G, Gao W (2007) Video2cartoon: generating 3d cartoon from video2cartoon: generating 3d cartoon from broadcast soccer video. IEEE Trans. Circuits Electron 53(3):1138–1146
9. Liu T, Yuan Z, Sun J, Wang J, Zheng N, Tang X (2010) Learning to detect a salient object. IEEE Trans Pattern Anal Mach Intell 33(2):1–15
10. Lai J-H, Chien S-Y (2009) Tennis video with semantic scalability. In: Proceedings of 11th IEEE International Symposium on Multimedia, pp 523–526
11. Recommendation JPEG Standard (1993) ITU-T
12. Recommendation H.264 (2003) Advanced video coding for generic audiovisual services. ITU-T
13. Tang Q, Koprinska I, Jin JS (2005) 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, pp 271–274
14. Thang TC, Kim J-G, Kang JW, Yoo J-J (2009) Svc adaptation: standard tools and supporting methods. Elsevier Signal Process. Image Commun 24(3):214–228
15. Wang Y, van der Schaar M, Chang S-F, Loui A (2005) Classification-based multidimensional adaptation prediction for scalable video coding using subjective quality evaluation. IEEE Trans Circuits Syst Video Technol 15(10):1270–1279
16. Wikstrand G, Eriksson S (2002) Football animations for mobile phones. In: Proceedings of the Second Nordic Conference on Human-Computer Interaction NordiCHI '02, pp 255–258
17. Yu X, Xu C, Leong HW, Tian Q, Tang Q, Wan KW (2003) Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video. In: Proceedings of the Eleventh ACM International Conference on Multimedia, pp 11–20

Jui-Hsin Lai received the B.S. degree in electronics engineering from the National Chiao-Tung University, Hsinchu, Taiwan, in 2005. In 2006, he is currently working toward the Ph.D. degree in the Media IC and System Laboratory, Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan. His research interests include the applications of sports video, interactive multimedia, video/image processing, and VLSI architecture design of multimedia processing.

Author's personal copy

Shao-Yi Chien (S'99–M'04) received the B.S. and Ph.D. degrees from the Department of Electrical Engineering, National Taiwan University (NTU), Taipei, Taiwan, in 1999 and 2003, respectively. During 2003 to 2004, he was a research staff in Quanta Research Institute, Tao Yuan County, Taiwan. In 2004, he joined the Graduate Institute of Electronics Engineering and Department of Electrical Engineering, National Taiwan University, as an Assistant Professor. Since 2008, he has been an Associate Professor. His research interests include video segmentation algorithm, intelligent video coding technology, perceptual coding technology, image processing for digital still cameras and display devices, computer graphics, and the associated VLSI and processor architectures. He has published more than 120 papers in these areas.

Dr. Chien serves as an Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology and Springer Circuits, Systems and Signal Processing (CSSP). He also served as a Guest Editor for Springer Journal of Signal Processing Systems in 2008. He also serves on the technical program committees of several conferences, such as ISCAS, A-SSCC, and VLSI-DAT.