REAL TIME SEMANTIC ADAPTATION OF SPORTS VIDEO WITH USER-CENTRED PERFORMANCE ANALYSIS

M. Bertini, A. Del Bimbo

D.S.I. - Università di Firenze - Italy
bertini,delbimbo@dsi.unifi.it

A. Prati, R. Cucchiara

D.I.I. - Università di Modena e Reggio Emilia - Italy
prati.andrea,cucchiara.rita@unimo.it

ABSTRACT

Semantic video adaptation improves traditional adaptation by taking into account the degree of relevance of the different portions of the content. It employs solutions to detect the significant parts of the video and applies different compression ratios to elements that have different importance. Performance of semantic adaptation heavily depends on the quality and precision of the automatic annotation, whether it operates in strict or non-strict real time, and the codec which is used to perform adaptation at the event or object level. It should consider the effects of the errors in the automatic extraction of objects and events over the operation of the adaptation subsystem, and relate these effects to the preferences for the objects and events of the video program, that have been decided by the user. In this paper, we present strict real time annotation and adaptation of sports video and introduce two new performance measures: Viewing Quality Loss and Bit-rate Cost Increase, that are obtained from classical PSNR and Bit Ratio, but relate the results of semantic adaptation with the user's preferences and expectations.

1. INTRODUCTION

New tools to support Universal Multimedia Access must deal with the ever-increasing demand for large amounts of multimedia data (in particular, video) and with the heterogeneity of clients' devices. With current network constraints and device limitations, access to long sequences live video is obtained with poor fidelity, often unsatisfactory for the user, and with high costs. A possible solution is the *content-based (or semantic) adaptation* that optimizes the available resources by improving the quality of those parts of the video that are more interesting for the user.

In fact, syntactic adaptation, achieved via temporal or spatial downscaling, or bitrate reduction, is not yet fully satisfactory in terms of the cost/quality trade-off for users' mobile terminals with limited resources. One possible solution might be that of letting customers choose which events and objects they are most interested in, thus adapting the original video stream in a way that maintains the highest video quality for important events and objects, while reducing the quality of other events and objects, or not even coding them. Using this *semantic adaptation*, for instance in the transmission of a soccer game video, we might send good quality video only for the frames in which a goal is scored, or within the individual frames, providing high resolution only for the most relevant parts (e.g. those surrounding the players).

The annotation system should perform real time analysis of an incoming video stream, marking the beginning of a possibly interesting sequence (containing an highlight) and then signaling the end of the interesting sequence. Possibly the system should be capable to even forecast a highlight (e.g. to provide real-time services such as mobile phone access) that will last some seconds with a certain probability. The probability may be used by the final user in order to select only certain highlights that are forecast with a minimum probability.

Since we are addressing a system that forecasts an highlight all the processing has to be performed in real-time. To this end we have considered a set of cues and a system architecture that allows to perform RT processing. A method to extract rapidly visual cues is to use features extracted from the compressed domain, e.g. MPEG motion vectors and MPEG DC components of DCT blocks.

This paper proposes an integrated framework of automatic annotation and semantic adaptation for live sports videos. Performance of semantic adaptation heavily depends on the quality and precision of the automatic annotation, whether it operates in strict or non-strict real time, and the codec which is used to perform adaptation at the event or object level. It should consider the effects of the errors in the automatic extraction of objects and events over the operation of the adaptation subsystem, and relate these effects to the preferences for the objects and events of the video program, that have been decided by the user.

To evaluate this we propose two performance measures useful for semantic adaptation: *Viewing Quality Loss* and *Bit-rate Cost Increase*, that are obtained as suitable modifications from PSNR and Bit Rate. They relate the results of semantic adaptation to the errors in the automatic extraction of objects and events and the user's preferences and expectations.

2. PREVIOUS WORK

Automatic sports video annotation has been addressed by several authors, with increasing attention in the very recent years. In [7] Bayes networks have been used to model and classify American football plays using trajectories of players and ball. However, trajectories are entered manually, and not automatically extracted from the video stream. Kijak et al. [8] have used multimodal features to analyze tennis sports video structure. Models are used to integrate audio and visual features and perform stochastic modelling. Visual cues are used to identify the court views. Ball hits, silence, applause and speech help to identify specific events like scores, reserves, new serves, aces, serves and returns. Annotation of soccer videos has been addressed by a large number of researchers. In [11] MPEG motion vectors are used to detect events. In particular, they exploit the fact that fast imaged camera motion is observed in correspondence of typical soccer events, such as

shot on goal or free kick. Recognition of relevant soccer highlights (free kicks, corner kicks, and penalty kicks) has been presented in [1]. Low level features like the playfield shape, camera motion and players' position are extracted and Hidden Markov Models are used to discriminate between the three highlights. More recently, in [6], Ekin et al. have performed event detection in soccer video using both shot sequence analysis and visual cues. In particular, they assume that the presence of highlights can be inferred from the occurrence of one or several slow motion shots and from the presence of shots where the referee and/or the goal post is framed. In [2] a system based on FSMs, that detects several different soccer highlights such as shot on goal, placed kicks, forward launches and turnovers, using visual cues has been presented. Ball trajectory is used by Yu et al. [18]. In order to detect the basic actions and compute ball possession by each team. Kalman filter is used to check whether a detected trajectory can be recognized as a ball trajectory. Experiments report detection of basic actions like touching and passing. Examples of detection of basic highlights in volleyball, tennis and soccer are reported. [13] has reported on detection of Formula 1 highlights using a multimodal fusion of cues and dynamic Bayes networks.

Semantic adaptation can be done at different levels: in general, *semantic video transcoding* means a coding change based on video content [12]; *video abstraction* means the extraction of single entities such as key-frames or objects in the frame that are significant for the video [9]; *video summarization*, on the other hand, refers to the process of grouping together only meaningful temporal segments of the video, in order to send the user only the interesting parts or the highlights. Many systems with content-based adaptation have been proposed: IBM's Video Semantic Summarization Systems, described in [14], exploits MPEG-7 for semantic transcoding and summarization. Semantic annotation is provided manually by human experts; the user specifies his/her requests in terms of topic preference, topic ranking, query keywords, and time constraints, and the system delivers a video summary. In [16], Vetro et al. present an object-based transcoding framework that uses dynamic programming or meta-data, for the allocation of bits among the multiple objects in view. Applications of object- and event-based adaptation have been proposed in [15] and [4] in the context of video surveillance, where the frames of the original video stream are preliminarily segmented into regions and interesting video entities are detected and transmitted in high quality, while non interesting entities or background are sent in low quality. In [3], Chang et al. have performed real time detection of the salient events of a sport game, like baseball *pitching* or tennis *serving*, and accomplish event-based adaptation and summarization.

To compare different semantic transcoding systems each other, some authors have redefined the typical measures of performance: *Bit Rate* (BR) - that represents the required bandwidth or the absolute number of transferred bits per second -, *Peak Signal-to-Noise Ratio* (PSNR) and *Mean Squared Error* (MSE) - that measure video quality, in order to take into account the physical attitudes of the users and their expectations or preferences. The measures of the effects of the frequency distortion and the additive noise that affect the human perception system have been taken into account in [5]. In [17] a new approach has been proposed that overcomes the weaknesses of MSE and PSNR, where the image degradation is accounted to the image structural distortions and the image quality is referred to the perceptive satisfaction of the user. In [10], a *utility function* has been defined that makes explicit the relationships between resources (bandwidth, display, etc.) and utilities (objective or subjective quality, user's satisfaction, etc.).

3. REAL-TIME ANNOTATION

MPEG videos are used in order to extract as much visual features as possible from the compressed domain, to speed up the processing. In particular the system has been tested using MPEG-1 and MPEG-2 videos. A probabilistic framework (Bayesian networks) is used to detect interesting highlights, associating a confidence number to the beginning and end of sequences that may contain a highlight, and thus allowing end users to set a sensitivity threshold to the system. In fact in the envisioned use case some users may prefer to get only very probable highlights, e.g. to reduce the costs related to video transmission, while other users may prefer to see more actions, accepting false alarms.

Only visual features are used by the system, since audio features may not be always available. The features may be divided in two groups: *compressed domain features*, that are extracted directly from the MPEG video stream:

- Motion vectors: MPEG motion vectors are used to calculate indexes of camera pan and tilt, and an index of motion intensity
- Playfield: YUV color components are used to extract and evaluate the playfield framed.

and uncompressed domain features, that are extracted from images

- Players: players are extracted using previous knowledge of team colors (to improve precision) from uncompressed I frame: the ratio of pixels of the two teams is the cue used by the Bayes networks.
- Playfield lines: playfield lines are extracted from the uncompressed I frame: they are filtered out based on length and orientation.

The ratio of playfield framed allows to classify frames in three types: long, medium and close shot. The playfield area framed is classified in three zones, using the histogram of line orientation: left, center and right.

Evidence and inference are computed for each MPEG GOP (12 frames, i.e every 1/2 second in PAL video standard). If the highlight is predicted in the following 6 seconds (12 GOPs) the video is processed by the Bayesian validation networks. Conditional probabilities are updated every 2 secs. Four networks are used to predict highlights: two networks to predict attack actions (left-right) and two networks to predict placed kicks (left-right).

The system is able to detect if the predicted action is concluded by a shot on goal. In fact when there is a shot on goal typically there is a sequence composed by 3 phases: 1) Fast panning of main camera towards goal post (Long Shot); 2) Zooming on the player who kicked the ball (Medium Shot or Close Shot); 3) View of the crowd or close up of the trainer (Close Shot). To detect the shots on goal two networks are used, one for the left, and one for the right side of the playfield. The networks have the same structure, while the conditional probability tables of the nodes change every 2 seconds following the three typical phases described before.

4. THE TRANSCODING SYSTEM

In order to let users specify their interests in terms of objects and events we define a set C of *classes of relevance* which groups together the parts of the video that are of the same degree of interest to the user.

Specifically, a class of relevance groups entities of the ontology (namely objects and events) with the same degree of relevance for the user. Formally, each element of the set $C = \{C_1, ..., C_{N_{CL}}\}$ is defined as:

$$C_i =< \mathbf{o_i}, \mathbf{e_i} > \quad with \quad \mathbf{o_i} \subseteq O, \mathbf{e_i} \subseteq E \tag{1}$$

The relevance associated to each class is quantified by means of a weight assigned by the user. The semantic annotation engine extracts from the raw video the meaningful objects (o_i) and the events (e_i). Then, objects and events are assigned to their class of relevance C_i. We modified the codec supplied by the MPEG Software Simulation Group open source library to create an output video compatible with MPEG-2 viewers. A different quantization scale QS (ranging from 1 to 31) can be selected for each macroblock of the frame. The QS is used as a multiplier for the quantization matrix of the DCT: the higher the scale, the more compressed the macroblock will result; its value is assigned by the transcoding system according to the object/event detection and to user's preferences.

Since objects and events are aggregated into distinct classes of relevance, the classification errors in the annotation eventually reflect into a different compression than expected, and consequently into a loss of satisfaction of the user. In particular:

- Events and objects that are under-estimated (E_u, O_u) or missed (E_m, O_m) have a negative impact on the viewing quality since they are more compressed. The transmission costs paid by the user are instead lowered.
- Events and objects that are over-estimated (E_o, O_o) or falsely detected (E_f, O_f) affect negatively the transmission costs. In fact, non interesting parts that are classified as relevant are produced at a higher viewing quality, and will have transmission costs higher than expected.

Pixels of frames will form sets according to how they have been classified by the annotation system. It is then possible to derive new indices of Visual Quality Loss and Bit-rate Cost Increase, based on the definitions of PSNR and BR. The first effect results from over-compression due to under-estimation and miss conditions occurred in the annotation; the second effect results from higher Bit-rate due to over-estimations and false detections.

In particular, for each frame I^t, the viewing quality loss (VQL^t) is defined as 1 minus the ratio between the PSNR calculated over the set Err^t_Q (the set of under-estimated pixels) and the PSNR measured over the same set of pixels in the ideal case of no annotation errors:

$$VQL^t = 1 - \frac{PSNR_{Err^t_Q}}{PSNR^{ID}_{Err^t_Q}}. \tag{2}$$

Similarly, the bit-rate cost increase for each frame I^t, BCI^t, is defined as 1 minus the ratio between the requested BR in the ideal case of no annotation errors and the one requested in the real case, both calculated over the set of pixels Err^t_C (the set of over-estimated pixels).

$$BCI^t = 1 - \frac{BR^{ID}_{Err^t_C}}{BR_{Err^t_C}} \tag{3}$$

Viewing quality loss (VQL) and bit-rate cost increase (BCI) for a video clip are directly obtained from the definitions above, by averaging VQL^t and BCI^t:

$$VQL = \frac{\sum_{t=0}^{N} VQL^t}{N} \quad ; \quad BCI = \frac{\sum_{t=0}^{N'} BCI^t}{N'}, \tag{4}$$

where N is the number of the frames associated with the events of the clip that are taken into consideration, and N' is N plus the number of the frames of the events that have been falsely detected.

5. EXPERIMENTAL RESULTS

The video stream used for the test set are MPEG-1 and MPEG-2 videos at 25 frames per second (PAL standard) and with a resolution that is respectively of 360×288 and 720×576. The GOP length is 12 frames. 268 case examples ($\sim$ 90 min) collected from World Championship 2002 and European Championship 2004 have been used to test the annotation system.

- 172 highlights that have been concluded with a shot on goal (SOG): 134 attack actions (AA) and 38 Placed kicks (PK)
- 54 highlights that have not been concluded with a shot on goal (NSOG): 51 attack actions and 3 Placed kicks
- 42 Other Actions (OA)

Table 1 and 2 report precision and recall figures, and a breakdown of the classification of SOG, NSOG and OA actions, and attack actions and placed kicks. The average number of frames between the prediction and the appearance of a SOG action is 74,2 ($\sim$ 3 sec. for a PAL system). The lower precision of placed kick detection is due to cases of free kicks that are quite far from the goal box area; in this case the area is framed after kicking the ball.

Highlight type	Predicted and SOG recognized	Predicted and SOG not recog.	Not predicted	*Precision*	*Recall*
SOG	**151/172**	13/172	8/172	0,96	0,88
NSOG	7/54	**43/54**	4/54	0,74	0,80
OA	0/42	2/42	**40/42**	0,77	0,95
			Avg.	0,83	0,88

Expected behaviour
Acceptable results
Bad results

Table 1. Annotation performance of SOG, NSOG and OA

To evaluate the efficacy of the integrated approach, an example of the VQL and BCI resulted on a sample video is reported in Fig. 1. From this graph it is possible to notice two distinct peaks: the leftmost corresponds to a single frame in which the annotation system detects an interesting event when it is not actually present,

Highlight type	Correctly detected	Misclassified/ missed	*Precision*	*Recall*
AA	163/185	22/185	0,98	0,88
PK	37/41	4/41	0,63	0,91
		Avg.	0,81	0,895

Table 2. Annotation performance of attack actions and placed kicks

thus wasting bandwidth; the rightmost peak, instead, is due to the missing of the last part of the highlight, for which the corresponding quality is not sufficient. It is also worth noting that there is also a little waste of bandwidth in the missed frames corresponding to the right peak: this effect is probably due to the not aligned change of the quality with respect of the GOP delimitation that results in partial mismatch of the references in B frames.

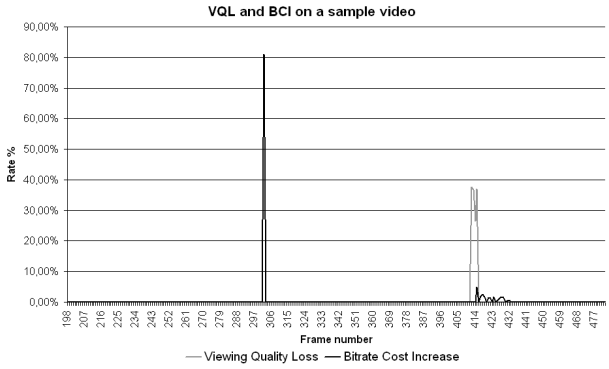

Fig. 1. Example of Viewing Quality Loss and Bitrate Cost Increase on a sample video.

6. CONCLUSIONS

In this paper we have reported the results of real time annotation and adaptation system, applied to soccer videos, that forecast the appearance of highlights in real-time, and transcode the video according to user preferences. Two new performance measures that take into account errors of the annotation system and user preferences have been introduced. Our future work will deal with a refinement of the proposed system, extending and specializing the types of highlights that may be forecast, and extending the system to other types of sports.

Acknowledgments

This work has been partially funded by the European VI FP, Network of Excellence DELOS (2004-06). Authors would like to thank Filippo Conforti for his help.

7. REFERENCES

[1] J. Assfalg, M. Bertini, A. D. Bimbo, W. Nunziati, and P. Pala. Soccer highlights detection and recognition using hmms. In *Proc. of Int'l Conf. on Multimedia and Expo (ICME2002)*, 2002.

[2] J. Assfalg, M. Bertini, C. Colombo, A. D. Bimbo, and W. Nunziati. Semantic annotation of soccer videos: automatic highlights identification. *Computer Vision and Image Understanding*, 92(2-3):285–305, November-December 2003.

[3] S.-F. Chang, D. Zhong, and R. Kuma;. Real-time content-based adaptive streaming of sports video. Technical Report 121, Columbia University, July 2001.

[4] R. Cucchiara, C. Grana, and A. Prati. Semantic video transcoding using classes of relevance. *International Journal of Image and Graphics*, 3(1):145–169, Jan. 2003.

[5] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik. Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9(4):636–650, Apr. 2000.

[6] A. Ekin, A. M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Transactions on Image Processing*, 12(7):796–807, July 2003.

[7] S. Intille and A. Bobick. Recognizing planned, multi-person action. *Computer Vision and Image Understanding*, 81(3):414–445, March 2001.

[8] E. Kijak, G. Gravier, L. Oisel, and P. Gros. Audiovisual integration for tennis broadcast structuring. In *CBMI 2003*, pages 421–428, Rennes France, 2003.

[9] C. Kim and J. N. Hwang. Object-based video abstraction for video surveillance systems. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(12):1128–1138, 2002.

[10] J.-G. Kim, Y. Wang, and S.-F. Chang;. Content-adaptive utility-based video adaptation. In *Proc. of IEEE Int'l Conference on Multimedia & Expo*, pages 281–284, July 2003.

[11] R. Leonardi and P. Migliorati. Semantic indexing of multimedia documents. *IEEE Multimedia*, 9(2):44–51, April-June 2002.

[12] K. Nagao, Y. Shirai, and K. Squire. Semantic annotation and transcoding: Making web content more accessible. *IEEE Multimedia*, 8(2):69–81, April-June 2001.

[13] M. Petkovic, V. Mihajlovic, and W. Jonker. Multi-modal extraction of highlights from tv formula 1 programs. In *Proc. of IEEE Int'l Conference on Multimedia & Expo*, 2002.

[14] I. research. *http://www.research.ibm.com/MediaStar/VideoSystem.html*.

[15] A. Vetro, T. Haga, K. Sumi, and H. Sun;. Object-based coding for long-term archive of surveillance video. In *Proc. of IEEE Int'l Conference on Multimedia & Expo*, volume 2, pages 417–420, 2003.

[16] A. Vetro, H. Sun, and Y. Wang. Object-based transcoding for adaptable video content delivery. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(3):387–401, March 2001.

[17] Z. Wang, A. C. Bovik, and L. Lu. Why is the image assessment so difficult? In *Proc. of the IEEE Conference on Acoustics Speech and Signal processing*, May 2002.

[18] X.Yu, C. Xu, H. Leung, Q. Tian, Q. Tang, and K. W. Wan. Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video. In *ACM Multimedia 2003*, volume 3, pages 11–20, Berkeley, CA (USA), 4-6 Nov. 2003 2003.