SEMANTIC ANNOTATION AND TRANSCODING FOR SPORT VIDEOS

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

Telecommunication companies are demonstrating interest in providing mobile video services. The availability of larger bandwidth, and the improvements in terms of resolution of the displays of third generation mobile phones, let telecom and content provider companies to provide new services to their customers. Among these services users can watch a certain number of sport videos, usually a selection of the best actions occurred during a play. In order to provide a timely and satisfying service to customers there is need of tools and systems that help to detect and recognize the interesting events, and optimize the use of bandwidth, coding these events and the most interesting objects within them at the best visual quality/bandwidth ratio.

1. INTRODUCTION

This paper addresses the problem of semantic annotation and transcoding of sport videos; this requires detection and recognition of sport highlights in videos, taking into account their temporal extension and detection of the players and objects that take part into the action. Results of this annotation drive a semantic transcoding system that adapts videos to the user's requirements and terminal constraints. This kind of research is motivated by the strong interest shown by telecommunication and mobile phone companies, who are interested in systems that ease the process of annotation and transcoding of sport videos that are provided to subscribers of video mobile services, that use third generation phones. Even if bandwidth availability is dramatically increased, if compared to former mobile phone systems, it is still limited, and a compromise between video quality and compression that fits the needs of each user has to be found.

A solution may be that of letting each customer to choose which events and objects are more interesting for her/him, thus transcoding the original video stream in a way that maintains the highest video quality that is possible for important events and objects, and reducing the quality of other events and objects, or even not coding them. Adapting videos to the user's requirements and terminal constraints is commonly referred to as *transcoding*. In particular semantic transcoding requires a video transcoding where the code change is driven by video content [5]. With *semantic transcoding*, the most meaningful parts of the video may have different coding than others; for example, in the transmission of a video of a soccer game, we can send good quality video only for the frames where interesting actions take place, or within the individual frames, provide high resolution sampling only for the most relevant parts (e.g. those in the surrounding of the players).

Research in semantic transcoding mostly concentrated on the extraction and separate coding of meaningful objects rather than of meaningful events with both spatial and temporal extension. Smith et al. in [8] proposed image analysis processes for content-based image transcoding using image type (e.g. graphs or photos) and image purpose classes. The IBM's Video Semantic Summarization Systems described in [6] exploits MPEG-7 for semantic transcoding: semantic annotation is provided manually by human experts; the user specifies his/her request in terms of preference topics, topic ranking, query keywords, and time constraint; the system outputs a video summary. In [5], Nagao et al. employ a video annotation editor that is capable of scene change detection, speech recognition, and correlation of scenes with the text obtained from the speech recognition engine. In this way, semantic indexes for video-to-document or video translation and summarization are produced. In [11], Vetro et al. presented an object-based transcoding framework that uses dynamic programming or meta-data, for the allocation of bits among the multiple objects in the scene.

In order to reach the goal of semantic transcoding an automatic annotation system of sport videos must be able to detect the beginning and the end of an highlight, and should also recognize some objects such as players, playfield or crowd. Automatic semantic annotation of sports video requires that the domain knowledge is properly included and exploited in the annotation process and that low and intermediate-level features are conveniently selected, extracted from the video and combined so that their spatio-temporal combinations identify the important highlights. Spatial and temporal extensions of the highlights must be precisely detected in order to permit the selection of the most salient parts of the video for transcoding, and video objects must be selected to code differently the interesting parts within a frame.

A number of researchers have provided evidence of the possibility of performing automatic annotation of semantic cues in sports video. Typical events of tennis have been modelled and detected in [10] using tennis court lines detection and player tracking. Rule-based modelling of complex basketball plays is presented in [12]. The knowledge base is represented as a decision-tree. Detection of events is performed by checking the occurrence rule predicates utilizing a visual low-level feature along with a threshold determined through training. In [9], shots of basketball game are classified into one of three categories using text detection, change in motion direction and detection of crowd cheering. Basket events are detected whenever the shot sequence displays certain audio/visual patterns. In [3], Ekin et al. performed highlight detection in soccer video using both shot sequence analysis and shot 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 box is framed. Detection of soccer highlights (free kicks, corners, and penalties) using Hidden Markov Models has been reported in [1]. Assfalg et al. [2] have presented automatic detection of the principal highlights in soccer, based on the estimation of a few visual cues. Each highlight has been mod-

elled with a Finite State Machine (FSM) and transitions from one state to the other are activated by the combination of visual cues extracted from the video stream.

In this paper we propose a framework for semantic annotation and transcoding of sport videos, based on automatic annotation of events and segmentation of objects extracted from an uncompressed video stream. We also propose a performance measure that combines quantitative measures of video quality and bandwidth, user preferences and satisfaction. Interesting highlights of a soccer match are modelled using FSMs. Transcoding is applied to the video stream according to the preferences of the user, that chooses the most interesting combinations of highlights and objects, that should be compressed maintaining high video quality.

The paper is structured as follows. The system framework and the metric for performance evaluation are presented in Section 2 and 3, respectively. The details on the algorithms used for annotation and transcoding are reported in Section 4. Experimental results are also presented. Conclusions are reported in Section 5.

2. THE PROPOSED FRAMEWORK

A *class of relevance (CoR)* is defined as the set of meaningful elements in which the user is interested in and that the system is able to manage. The importance of CoR is twofold. First, the set of classes defines an ontology of the scenario that must be recognized, annotated, and provided to the user. Secondly, the user can exploit the classes of relevance in order to define his/her preferences about the video content, thus driving the transcoding process in order to achieve the desired quality/cost trade-off. In addition, sets of classes can be used for performance evaluation purposes, as reported in the next section. For our purposes, the set of classes of relevance includes all the *events* and *objects* of the scene that can be automatically identified and transcoded.

Formally, a *class of relevance* C is defined as a pair $C =< o_i, e_j >$, where o_i represents an object class and e_j is an event class, selected between the set of object classes O and event classes E detectable by the system:

$$O = \{o_1, o_2, ..., o_n\} \cup \{\widetilde{o}\} \quad ; \quad E = \{e_1, e_2, ..., e_m\} \cup \{\widetilde{e}\}$$

The special class $\widetilde{o}$ includes all the areas of the image that do not belong to user-defined classes (for example, the part outside the soccer playfield can be considered as $\widetilde{o}$). Analogously, the event $\widetilde{e}$ includes all the non interesting events or the case of no-event.

As an example, let us define the set O and E in the case of soccer videos. A possible set of objects that the system is able to segment is represented by:

$$O = \{PF, PU, PL\} \cup \{\widetilde{o}\}; E = \{SG, PK, TO, FL\} \cup \{\widetilde{e}\} \quad (1)$$

where PF, PU, and PL stay for "playfield", "public" (i.e., out of the playfield), and "players", respectively. In this case PF, PU, and PL are considered as a complete partition of the image, i.e. $\widetilde{o}$ is null. For the events, SG, PK, TO and FL correspond to the followings: "shot at goal", "penalty kick", "turnover" and "forward launch". We can define the set of *classes of relevance* $C = \{C_i\}$ as the set of all the feasible combinations between each object and each event. Among the classes of relevance the user can select a set UC that has two characteristics: *i*) its elements are selected among the elements of C; *ii*) the user can group different C_i in a single element of UC.

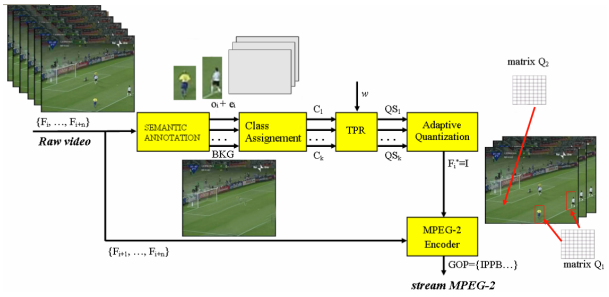

Fig. 1. Scheme of the semantic annotation and transcoding system used.

The scheme reported in Fig. 1 displays the process of semantic transcoding adopted in our system, called SAQ-MPEG. 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. The TPR engine (Transcoding Policy Resolver) computes the quantization multipliers QS_i according to the user's defined relevance weights assigned to the UC set to which the class belongs. By multiplying the QS_i with the MPEG-2 quantization matrix a quantization matrix for each class of relevance is obtained. Finally, a standard MPEG-2 encoder uses this coded frame as I frame and creates the GOP (Group Of Pictures) of the stream. The MPEG-2 standard has been chosen because of its capabilities of temporal prediction, to reduce the required bandwidth and produce a video that can be played by a standard decoder.

Automatic annotation performs the extraction of low-level features and their classification by means of high-level modules that are tailored on the specific application. For example, we partition the soccer playfield into a number of different zones with slight overlapping and use the motion of the main camera as a cue for the description of the evolution of the play. Each event is modeled with a Finite State Machine, where key actions, defined in terms of the estimated cues, determine the transition from one state to the following. The event models are checked against the current observations, using a model checking algorithm. The objects of interest extracted are the playfield zones, the player blobs and the background. A short description of this subsystem is reported in section 4, while interested readers may consult the detailed description provided in [2].

The semantic extracted is used to drive the *adaptive quantization* of frame I in the MPEG stream (see Fig. 1). This results into standard MPEG stream, but with different compression within the frame, according to the image region that is under examination.

3. PERFORMANCE EVALUATION METRIC

Performance evaluation of annotation and transcoding systems is typically based on a comparison with ground-truthed data obtained from manual annotation. In the case of annotation, comparison is made at object- or event-level by collecting errors or computing a confusion matrix with false positives and negatives. Instead, in the case of transcoding, the comparison is usually at pixel-level by computing figures, such as the PSNR (Peak Signal-to-Noise Ratio), that evaluate the difference between original and distorted (adapted) images.

We propose a performance evaluation that takes into account *user satisfaction* in terms of perceived visual quality and in terms of costs. The basic visual quality measure used is MSE (mean

square error) of the coded pixels; in the considered use case (i.e, mobile communications), the cost is typically associated with the bandwidth, and thus can be measured in terms of bitrate (BR).

For each highlight requested by the user in a class of relevance the annotation system ma have the following behaviours: *i)* correct detection (EA); *ii)* missed detection (EB); *iii)* false detection: the detected highlight may have a higher ($EC1$) or a lower ($EC2$) relevance than the actual event. The same effects are possible for object detection (OA, OB, $OC1$, $OC2$). The effects of these errors are: *i)* bandwidth waste, that the user may have to pay; *ii)* viewing quality loss; e.g. $EC1$ leads to bandwidth waste, EB or OB lead to loss of viewing quality. From these considerations we may derive a set of user satisfaction levels, starting from a perfect annotation (USA=$EA \wedge OA$), and then taking into account errors on object detection ($USB1 = EA \wedge (OB \vee OC2); USB2 = EA \wedge (OC1)$, and finally errors in highlight detection ($USC1 = EC2 \wedge (OB \vee OC1 \vee OC2); USC2 = EC1 \wedge (OB \vee OC1 \vee OC2); USDD = EB2 \wedge (OB \vee OC1 \vee OC2)$). According to the effect on signal degradation or bandwidth increase we can define the following quantities, related to the user satisfaction classes:

$$\alpha_1 = \frac{MSE_{USA}}{MSE_{USB_1}}; \alpha_2 = \frac{MSE_{USA}}{MSE_{USC_1}}$$
$$\beta_1 = \frac{BR_{USA}}{BR_{USB_2}}; \beta_2 = \frac{BR_{USA}}{BR_{USC_2}}; \beta_3 = \frac{BR_{USA}}{BR_{USD}} \quad (2)$$

The α measures can be related to a single pixel of each image, being the MSE_{USA} the Mean Square Error computed on the ideal transcoding result (USA), i.e. in the case there are no errors in the automatic detection subsystem (the pixel is assigned to the correct class of relevance). The signal distortion for the USA case is computed w.r.t. the original uncompressed frame, and is due to the lossy compression algorithm used. This distortion is unavoidable and is dependent on the class UC_i which the current pixel belongs to. On the other hand, the β measures can be defined only at frame-level, being BR the achieved bitrate, i.e. the bandwidth occupied by the frame.

We can define two separated values to measure the error committed in a frame, one in terms of viewing quality loss, the other in terms of bandwidth allocation. The former can be defined for each class UC_i

$$E^{frame}_{V_{UC_i}} = \frac{\sum_{p \in UC_i} (1 - \alpha^i_1(p)) + (1 - \alpha^i_2(p))}{|UC_i|} \quad (3)$$

where $\alpha^i_1(p)$ is, according with eq. 2, the α_1 defined for the class UC_i in the case of the point p. Obviously, if the point p does not belong to the user satisfaction class USB_1, the corresponding α_1 will be 1. Similar considerations can be done for $\alpha^i_2(p)$. Please, note that the α measures range between 1 (in the case the MSE of the pixel is that of the ideal transcoding USA) and 0 (in the case the ideal transcoding does not introduce any distortion). As a consequence, we use as normalization factor the number of pixel p belonging to the class UC_i. Thus, $E^{frame}_{V_{UC_i}}$ is equal to 0 in the case none of the pixel p belonging to the class UC_i are misclassified, and it is equal to 1 in the case all the pixel of the class have an infinite error due to the automatic detection subsystem.

The error in terms of bandwidth allocation, is defined only as measure at frame-level, being the bitrate difficult to be defined for a single pixel:

$$E^{frame}_{B_{UC_i}} = \sum_{j=1}^{3} (1 - \beta^i_j) \quad (4)$$

where, for example, β^i_1 is the β_1 measure of the current frame belonging to the class UC_i. In fact, the event associated to the current frame and the corresponding event in the ground-truth allow us to define in which of the user satisfaction classes this frame falls. If this class is USB_2, USC_2 or USD we can compute the corresponding β for that frame, by computing the bandwidth allocated for the current frame and comparing it with that allocated in the case of no annotation errors (USA). The β measures range between 1 (in the case of no errors) and $\frac{BR_{USA}}{BR_{uncompr}}$, that is, the bitrate is upper bounded by the bandwidth allocated by the uncompressed frame.

We can define a global frame-wise measure of the two types of errors as follows:

$$E^{frame}_V = \sum_{i=1}^{N_{CL}} w_i \cdot E^{frame}_{V_{UC_i}}; E^{frame}_B = \sum_{i=1}^{N_{CL}} w_i \cdot E^{frame}_{B_{UC_i}} \quad (5)$$

where w_i are the weights used by the user to measure the importance of the classes of relevance. These measures can be also summed up over the consecutive frames that belong to the detected event or over the complete video stream.

4. SEMANTIC ANNOTATION AND ADAPTATION OF SPORT VIDEOS

The system used to perform automatic annotation of sport videos is based on finite state machines and model checking. This choice appears to be a general and effective approach to model and detect highlights in sports video. Constraints for the finite state machine state transitions have been modeled with temporal logic. In the following we will report only on soccer to spare space.

We model highlights using finite state machines: each highlight is described by a directed graph $\mathcal{G}^h = \langle \mathcal{S}^h, \mathcal{E}^h \rangle$, where $\mathcal{S}^h$ is the set of nodes representing the *states* and $\mathcal{E}^h$ is the set of edges representing the *events*. Events indicate transitions from one state to the other: they capture the relevant steps in the progression of the play or of the race, such as moving from one part of the playfield to a different one, accelerating or decelerating, etc. State transitions are determined by different cues, that are directly estimated from visual data. Cues may be common to different sports. The playfield zone that is framed and camera motion are the cues that are used to describe and identify state transition conditions. Combinations of visual cues descriptors that determine state transitions are created through logic and relational operators. Time constraints (for example a minimum temporal duration) can be applied to some state transitions.

Soccer highlights that we have modeled with this approach are: *i*) forward launches, *ii*) shots on goal, *iii*) turnovers, *iv*) placed kicks (comprising penalty kicks, free kicks next to the goal box, and corner kicks). During the annotation processing the playfield, the players blobs and the crowd blobs are extracted from each frame (using color analysis, k-fill and morphological operations), and are used both for annotation and for transcoding purposes, since they are used to identify the O objects in the relevance classes.

Objects aura

To improve the visual appearance of the encoded objects, the blobs that identify them are enlarged, selecting an "interest aura" around

them. This aura is calculated taking into account the foveation effect of the HVS (human vision system); the HVS does not perceive an entire visual stimulus at full resolution because of non-uniform spacing of sensors ([7]): only objects that are comprised within the fovea area of the retina are perceived at high resolution. Thus around each border pixel of the blobs is selected a patch of pixels that fall within the radius of the foveola, that is the part of the retina with the highest density of sensors. This enlarged area provides a context for each object that enhances readability of the object itself, and eases the overall understanding of the action. Since calculation of this area requires knowledge of display characteristics and its distance from the eyes of the viewer, we have selected a Sharp Zaurus SL-C700 with a $3.7''$, 640×480 pixels display as target device for transcoding. From anthropometric measures ([4]) we have considered that it may be viewed from a distance of about 40 cm. The transcoded video has been scaled to fit the selected display, and the calculation of the object aura has been done using the above mentioned display characteristics and viewing distance.

Experiments

To evaluate the proposed metric several videos were manually segmented to obtain a ground truth both in terms of temporal extension of an highlight and in terms of objects, segmenting playfield, crowd and players. These videos were then processed by the automatic annotation system and the two types of user unsatisfaction, the one due to bandwidth waste and the other due to visual quality loss, have been calculated. An example of this analysis is reported in figure 2; we will report on experiments more thoroughly in the extended version of this paper. Within this sequence there are two highlights: a placed kick and a shot on goal. Two simple sets of CoR were used: $C_1 = \{< PF, SG >, < PF, PK >\}$ and C_2 = everything else; the weights are $\{w_h, w_l\} = \{0.7, 0.3\}$. According to the ground truth the start and end frames for these highlights are: 0 and 227 for the first highlight, and 262 - 302 for the second. The automatic annotation system detected these highlights with the first starting at frame 0 and ending at frame 232, while the second goes from 266 to 302. The effects of the misclassification of some frames is shown by the two spikes in the graph. The left spike is due to the frames that were classified as belonging to a "placed kick" highlight: this has lead to a low compression, to maintain the visual quality, and thus has resulted in a bandwidth loss. The right spike is due to the opposite effect: 4 frames were not classified as "shot on goal" and thus were more heavily compressed, resulting in visual quality loss. It can be noted that there is a small visual quality loss, due to misclassification of some pixels within the highlights. Figure 3 shows a comparison between a standard MPEG-2 encoding and the SAQ-MPEG algorithm, using PSNR. The bandwidth are: 1445,55 kbps for MPEG-2 and 1455,52 kbps for SAQ-MPEG. It can be noted that while maintaining the same bandwidth requirement SAQ-MPEG achieves a higher PSNR within the interesting highlights.

5. CONCLUSIONS

In this paper we have presented a framework for event-based and object-based semantic annotation for sport videos and semantic transcoding. The use case considered is that of mobile video services, but the system may be used whenever there are constraints on bandwidth, or there is need to take into account user preferences when compressing a video; in fact it is possible to select events and objects that deserve a higher visual quality, or to select

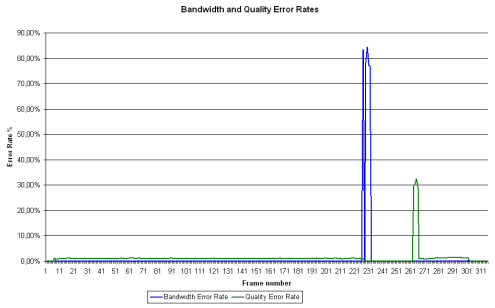

Fig. 2. User satisfaction error rate: bandwidth allocation error rate (left spike) due to classification of frames as having a higher relevance then the actual one, and viewing quality error rate (right spike) due to classification of high relevance frames as low relevance ones.

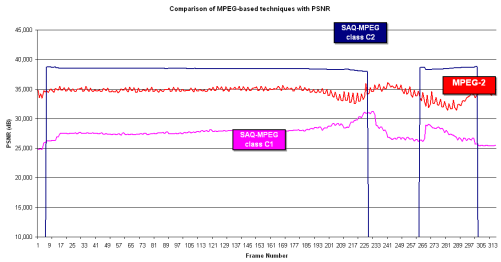

Fig. 3. Comparison of MPEG-based techniques with our SAQ-MPEG with standard PSNR. Bandwidth occupations are 1445,55 kbps for MPEG-2 and 1455,52 kbps for SAQ-MPEG total.

other events and objects that may be more heavily compressed. Since transcoding is driven by the annotation system the overall performance is dependent on the accuracy of annotation and segmentation. We have also introduced a performance measure that takes into account the effects due to annotation errors and user preferences.

6. REFERENCES

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

[4] S. Moss, Z. Wang, M. Salloum, M. Reed, M. van Ratingen, D. Cesari, R. Scherer, T. Uchimura, and M. Beusenberg. Anthropometry for worldsid a world-harmonized midsize male side impact crash dummy. Technical Report 2000-01-2202, SAE International, 2000.

[5] 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.

[6] IBM research. *http://www.research.ibm.com/MediaStar/VideoSystem.html*.

[7] H.R. Sheikh, B.L. Evans, and A.C. Bovik. Semantic indexing of multimedia documents. *Real-Time Imaging*, 9(1):27–40, February 2003.

[8] J.R. Smith, R. Mohan, and C. Li. Content-based transcoding of images in the internet. In *Proc. of IEEE Int'l Conference on Image Processing*, volume 3, pages 7–11, October 1998.

[9] S.Nepal, U.Srinivasan, and G.Reynolds. Automatic detection of 'goal' segments in basketball videos. In *Proc. of ACM Multimedia*, pages 261–269, 2001.

[10] G. Sudhir, J.C.M. Lee, and A.K. Jain. Automatic classification of tennis video for high-level content-based retrieval. In *Proc. of the Int'l Workshop on Content-Based Access of Image and Video Databases (CAIVD '98)*, 1998.

[11] 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.

[12] W. Zhou, A. Vellaikal, and C.C.J. Kuo. Rule-based video classification system for basketball video indexing. In *ACM Multimedia 2000 workshop*, pages 213–126, 2001.