Semantic Adaptation of Sport Videos With User-Centred Performance Analysis

Marco Bertini, Rita Cucchiara, *Member, IEEE*, Alberto Del Bimbo, *Member, IEEE*, and Andrea Prati, *Member, IEEE*

***Abstract*—In semantic video adaptation measures of performance must consider the impact of the errors in the automatic annotation over the adaptation in relationship with the preferences and expectations of the user. In this paper, we define two new performance measures *Viewing Quality Loss* and *Bit-rate Cost Increase*, that are obtained from classical peak signal-to-noise ration (PSNR) and bit rate, and relate the results of semantic adaptation to the errors in the annotation of events and objects and the user's preferences and expectations. We present and discuss results obtained with a system that performs automatic annotation of soccer sport video highlights and applies different coding strategies to different parts of the video according to their relative importance for the end user. With reference to this framework, we analyze how highlights' statistics and the errors of the annotation engine influence the performance of semantic adaptation and reflect into the quality of the video displayed at the user's client and the increase of transmission costs.**

***Index Terms*—Automatic annotation, object- and event-level compression, performance analysis, performance indices, semantic video adaptation, soccer sport video.**

I. Introduction

TRADITIONAL adaptation approaches transform the video presentation in the final format independently of its content. They perform video compression using the codecs available at the client, accomplish media scaling, for example by frame resizing, and adapt video transmission rate and presentation quality to the bandwidth available and the user's device [1]–[3]. Semantic or content-based adaptation, instead, alters the video format with respect not only to the transmission and presentation constraints of the end user, but also to the semantic content of the video. Video components are detected and classified into prioritized categories according to the user's preferences. The content of the presentation can be customized by temporal or spatial summarization (the content is reduced to a smaller duration or frame segments are removed), or abstracted (the video is transformed in a sequence of single entities like significant key-frames), or transformed into a different media (like audio or text). Different coding parameters can be applied to different parts of the video according to their relative importance for the end user. Semantic adaptation takes much of its appeal from the emergence of *Universal Multimedia Access*. In fact, mobile devices like PDAs or third generation cellular phones, while presenting a large range of display, storage and connection capabilities, should adapt to the personal expectations of the individual user in viewing media content. In particular, in the transmission of video streams, there is the need that adaptation is guided by both the operational constraints and the *events* and *objects* of the video that have some specific value for the user. In this way, the bandwidth is fully exploited for the transmission of the relevant parts of the video in high quality, while less interesting parts are transmitted in low quality, or transcoded into text, or definitely not transmitted, so as to minimize the usage of the resources available. The importance of these aspects is recognized in the MPEG-21 standard [4], that takes into account the identification of the terminal capabilities, the characteristics of the network, the definition of the user's profile (personal information, usage preferences and presentation preferences), and the physical environmental conditions around the user.

Automatic annotation of semantically meaningful parts of video content is a prerequisite for effective semantic video adaptation. Pattern recognition solutions can be employed to detect specific visual and auditory patterns that identify clips of relevant events or frame segments corresponding to meaningful objects. Integrated solutions that perform semantic adaptation with automatic annotation of meaningful parts or elements have been proposed by a few authors. Approaches that employ *attentive models* derive vague semantic indexes, like saliency points or high entropy regions, from low-level perceptive cues of the video stream [5]. In that low level cues are obtained from color and texture or motion vectors, with these approaches adaptation can be made in the compressed domain, avoiding decompression and re-compression. Adaptation in the compressed domain has been obtained by *requantization* [6], [7], *spatial resolution downscaling* [8], *temporal resolution downscaling* [9], or by a combination of them [10]. VideoEd [2], [11] is an example of semantic adaptation engine working at compressed level that performs requantization and spatial and temporal downscaling. Approaches that perform *adaptation based on objects and events* must operate in the noncompressed domain. Applications of object- and event-based adaptation have been proposed in [3] and [12] 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 and background are sent in low quality. In [2] and [13], the annotated video is stored in a database server

Manuscript received April 10, 2004; revised April 27, 2005. This work was supported in part by the Italian FIRB project (PERF-Performance analysis of complex systems 2003–05) and the European VI FP, Network of Excellence DELOS (2004–06) on digital libraries. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yoshinori Kuno.

M. Bertini and A. Del Bimbo are with Dipartimento di Sistemi e Informatica, Università di Firenze, Firenze 50139, Italy (e-mail: bertini@dsi.unifi.it; delbimbo@dsi.unifi.it).

R. Cucchiara and A. Prati are with Dipartimento di Ingegneria dell'Informazione, Università di Modena e Reggio Emilia, 41100 Modena, Italy (e-mail: cucchiara.rita@unimore.it; prati.andrea@unimore.it)..

Digital Object Identifier 10.1109/TMM.2006.870762

VideoAL, and is hence adapted according to the user's requests and device's capabilities, so as to create video summaries with the appropriate format, size and quality; annotations are high-level concepts like *house setting* or *man-made scene* that are obtained from either manual input or model learning classification. In [14], Nagao *et al.* have employed 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. For sport videos, most of the literature aims to perform highlight detection with the ultimate goal of supporting content-based video retrieval, and only a few authors have addressed the problem of automatic annotation and semantic adaptation for selective transmission and display of video content. Notably, in [15], Chang *et al.* have performed real time detection of the salient events of a sport game, like *race start* or *shot on goal*, from uncompressed video and accomplish event-based adaptation and summarization.

The applicability of semantic adaptation in real contexts requires that its operation is measured with appropriate performance indices. A few measures of performance have been derived that consider the effects induced by the application of different coding parameters to different parts of the video, or their relationships with user's physical attitudes, preferences or expectations. Effects of using object-based video coding on the improvement of *bit rate* (BR), *Peak signal-to-noise ratio* (PSNR), and *mean squared error* (MSE) have been discussed in [12], [16], [17], and [18]. In [19], BR and PSNR measures of performance have been redefined in order to take into account the effects of the frequency distortion and the additive noise that affect the human perception system. In [20], image degradation has been accounted to the image structural distortions, and image quality has been put in direct relationship with the perceptive satisfaction of the user. A *weighted PSNR* has been defined in [21], so as to take into account both the distortion due to the coding policy and the user's preferences. In [22], 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.).

However, realistic user-centred performance measures of semantic adaptation should also consider the effects of the errors in the automatic extraction of objects and events over the output of the adaptation, in relationship with user's preferences and expectations. In this paper, we define two new performance measures for semantic adaptation, namely *Viewing Quality Loss* and *Bit-rate Cost Increase*, that measure the effects of the errors in the annotation, when different coding parameters are applied to parts of the video that have different importance for the user. *viewing quality loss* and *bit-rate cost increase* are obtained from PSNR and bit rate, and measure respectively the loss of viewing quality and the bandwidth waste in reference to the user's expectations. User's preferences are expressed through the definition of *classes of relevance*, each of which assigns to a set of objects and events that have the same relevance, a relative relevance weight so that the adaptation engine provides different coding options for objects and events of different classes.

We discuss the critical factors that affect performance referring to a prototype system, that performs automatic annotation of meaningful highlights and objects of soccer video, and applies different coding parameters at either the event or the event and object level. Results are presented for a number of sample user profiles that represent typical end users.

The rest of the paper is organized as follows. In Section II, we discuss critical factors that affect performance of automatic annotation and semantic adaptation of soccer videos, considering the possible errors of the annotation engine and different implementations of the codec in the adaptation engine. In Sections III and IV, we introduce the new measures of performance and discuss the performance of the semantic adaptation engine in different cases. Conclusions are reported in Section V.

II. Semantic Annotation and Adaptation

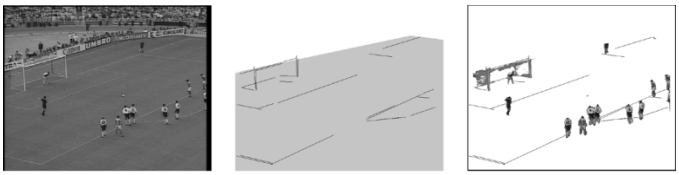
In the following, we present principles of operation and performance figures of automatic annotation of soccer video highlights. Next we discuss different implementations of video codecs based on MPEG2 and MPEG4, that are used to apply selective coding to the video stream performing content-based compression at the event and object-event level.

A. Automatic Annotation

Automatic annotation of soccer sport videos requires the identification of a limited number of visual cues, that are sufficient for the recognition of the most important events and entities. Under the assumption that a single main camera is employed to follow the action of the play, the prototype annotation engine implemented performs automatic annotation of soccer video principal highlights and entities, based on *camera motion*, *playfield zone*, and *players' position in the playfield*, that are detected in the video frames.

Camera motion is intimately related with the development of the play and can be used instead of other cues to provide affordable indications on the modes in which the play action develops (for example, camera motion tracking can replace ball motion tracking). The presence of multiple independent motions (like crowd's or players' motions) may affect negatively the estimation of camera motion. To overcome this problem, in our implementation, corners are extracted and tracked frame by frame and motion vectors are clustered following a deterministic sample consensus [23]. For each trajectory, the two nearest trajectories are used to compute the affine motion transformation (since the camera is in a fixed position, three parameters suffice to obtain a reasonable estimation of camera pan, tilt and zoom); multiple independent image motions are separated, by making each motion trajectory vote for the closest transformation. Camera motion is then obtained as the motion transformation with the highest consensus.

Playfield zones are useful to understand where the play takes place. Dividing the playfield into zones allows to break down a single action into a finite combination of phases, in each of which the play is characterized by a typical behavior (for example, a *forward launch* can be modeled as "slow motion in the central playfield zone followed by fast motion in the zone close to the goal post"). In our case, the soccer playfield has been divided into 12 distinct zones, six for each side, such that passing



(a) (b) (c)

Fig. 1. (a) Original frame; (b) playfield shape and playfield lines; (c) soccer players' blobs and playfield lines.

from one zone to the other corresponds to a new phase in the play. Each playfield zone, when viewed from the main camera, has a *typical view* that corresponds to a particular shape of the playfield region framed. Classification of the playfield view into one of the 12 playfield zones is based on: the shape of the playfield region; the area of the playfield region; the position of the region corner; the orientation of the playfield lines (the white lines in the soccer playground); and the midfield line position. The playfield region is segmented from color histogramming, by using grass color information. The shape of the playfield region is then refined by applying a processing chain composed of K-fill, flood fill, and the morphological operators of erosion and dilation. The playfield line segments are extracted from the edge map of the playfield region using a region growing algorithm [24]. The playfield lines are then obtained by joining together the white segments that are close and collinear. Twelve independent naive Bayes classifiers, one for each playfield zone, are used for classification. Each of them outputs the probability that the playfield region framed corresponds to the playfield zone of the classifier. Output probabilities are compared following a maximum likelihood approach [25]. A fixed-length temporal window and majority voting are used to filter out instantaneous misclassifications. The classification of playfield zones is reliable in most of the frames because, typically, the playfield region color is almost uniform and there are no large occlusions of playfield region lines determined by players' blobs.

Players' positions in the playfield are used in those cases where highlights with similar phases differ in the deployment of the players in the playfield. Typical cases are, for example, penalty kicks and free kicks. In our approach, blobs of individual players or groups of players are segmented out from the playfield by color differencing, and an adaptive template matching of an elliptical template, with height/width ratio equal to the median of the height/width ratios of the blobs extracted, is used to identify the players' silhouettes. The player's position on the playfield is computed from the projective transformation (the 3 $\times$ 3 matrix of the planar homography), which maps a generic imaged playfield point onto the real playfield point, and whose eight independent entries depend on both camera position and calibration parameters [26]. The planar homography entries are directly obtained using the correspondences in two subsequent frames of four lines, selected from the set of line segments obtained from playfield region segmentation. Fig. 1 shows a typical result of the extraction of the playfield region shape, playfield lines, and players' silhouettes.

Highlights that have been modeled are those typically reported in broadcasters' summaries of soccer matches, namely Forward launch, Shot on goal, Placed kicks, Attack action, and Counter attack. They are modeled with finite state machines, where nodes represent the action phases, and edges represent conditions over the visual cues under which a state transition occurs, eventually with constraints of temporal duration. Highlights are detected by using model checking. A detailed description of the finite state models and the algorithms that have been implemented is provided in [27].

TABLE I
PRECISION AND MISCLASSIFICATION RATES OF SOCCER HIGHLIGHT AUTOMATIC ANNOTATION

DETECTED HIGHLIGHT	EVENT CLIPS					
	Forward launch	Shot on goal	Placed kick	Attack act.	Counter att.	No highlight
Forward launch	**89.75%**	1.67%	0.00%	0.0%	0.00%	8.58%
Shot on goal	1.52%	**93.90%**	0.00%	0.00%	0.00%	4.58%
Placed kick	0.00%	0.00%	**89.75%**	0.00%	0.00%	10.25%
Attack action	1.50%	1.10%	0.00%	**96.40%**	1.00%	0.00%
Counter attack	0.00%	0.00%	0.00%	8.33%	**83.34%**	8.33%

TABLE II
MISS RATES OF SOCCER HIGHLIGHT AUTOMATIC ANNOTATION

EVENT CLIPS				
Forward launch	Shot on goal	Placed kick	Attack action	Counter attack
5.12%	13.33%	7.14%	25.00%	20.00%

TABLE III
AVERAGE NUMBER OF FRAMES THAT ARE MISSED OR FALSELY DETECTED FOR EACH SOCCER HIGHLIGHT, AT THE BEGINNING AND THE END OF THE EVENT

DETECTED HIGHLIGHT	Missed (begin)	False (begin)	Missed (end)	False (end)
Forward launch	9.51	0.03	2.84	0.43
Shot on goal	3.33	2.93	0.27	0
(Counter)Attack action	3.21	0	0.41	0.42
Placed kick	1.42	0.75	6.75	0.08

Table I reports precision and misclassification rates of highlight detection. Miss rates are shown in Table II. Attack action and Shot on goal have the lowest misclassification rate. Nevertheless they present high miss rates. Forward launch and Placed kick have low miss rates and relatively good precision figures. Counter attack detection has the worst performance figures.

Table III reports statistics of the errors at the frame level, for each type of highlight that is correctly detected. In particular, they are shown the number of highlight frames that are missed (i.e. the highlight detected is shorter than the real one) and the number of frames that are falsely detected (i.e., the highlight detected is longer than the real one), both at the beginning and at the end of the event. It is worth noticing that frame misses have higher incidence than falses. Forward launch has the highest frame miss rate at the beginning of the event; Placed kick the highest frame miss rate at the end. Shot on goal has the highest frame false detection rate at the beginning of the event. False detections at the end of the event are negligible for all highlights.

Results have been obtained testing the annotation engine over a test set of about 90' of standard PAL 720 $\times$ 576 video clips, extracted from soccer sport videos of the 2004 European championship games by Sky TV, and other national soccer games by BBC, RAI, and RTZ.

The relative importance of each highlight in soccer can be derived from Table IV. For each highlight, they are shown the average relative frequency of occurrence, as from UEFA statistics [28], and the highlight average duration, as observed for the

TABLE IV
AVERAGE HIGHLIGHT FREQUENCY FROM UEFA STATISTICS AND AVERAGE ESTIMATED HIGHLIGHT DURATION (NUMBER OF FRAMES)

HIGHLIGHT	UEFA AVERAGE FREQUENCY	AVERAGE DURATION OBSERVED
Forward launch	42.7%	37
Shot on goal	24.4%	62
(Counter)Attack action	14.6%	75
Placed kick	18.3%	112

TABLE V
AVERAGE FALSE DETECTION AND MISS RATE PER FRAME OF THE PIXELS OF RELEVANT OBJECTS IN SOCCER

OBJECT	False detection rate	Miss rate
Playfield	1.60%	4.50%
Players	5.32%	6.30%

test set used. The average number of Forward launch in a typical soccer game is almost twice as the number of Shot on goal and Attack action. The average number of Placed kick is about two thirds of Shot on goal. Placed kick has an average duration that is almost twice as Shot on goal and three times as Forward launch.

Detection and classification of Playfield and Player objects is typically very robust. Table V indicates the average false detection and miss rates per frame, of the pixels of playfield and player objects, as observed over the test set.

In that annotation errors reflect on the overall performance of semantic adaptation, performance figures of the annotation engine and the indications of the frequency of occurrence and average duration of the principal highlights are important factors to be considered. They can be used to guide the design of the user's preferences and to derive indications on the performance that can be expected for a specific user profile, as discussed in detail in Section III.

B. Content-Based Adaptation

Distinct solutions of content-based adaptation have been implemented at the event-level and at the object and object-event level, based on both MPEG2 and MPEG4 codecs. Adaptation at the event level applies different compression rates to the highlights and the rest of the video, according to their relevance, as assigned by the users; adaptation at the object-event level applies selective compression to objects only for the duration of the event. For testing, we have used the same test set as used for the test of the annotation engine. The soccer events have been distinguished into three distinct classes, of decreasing relevance: Shot on goal and Placed kick, Forward launch, and the other nonhighlights events. We have then applied DCT quantization scales of 5, 20, 31, (being 1 and 31 associated with the best and worst quality, respectively) to the events in the three classes, respectively. PAL video frames have been downscaled to the $3.7''$, 640×480 pixels (85.1 pixel/cm) of the Sharp Zaurus SL-C700 display. The performance and the average improvement in bandwidth allocation and PSNR, obtained with content-based adaptation with respect to standard coding, are briefly summarized in the following, separately for MPEG2 and MPEG4 based coding. A detailed discussion of the codec implementations has been published in [29].

1) MPEG-2 Based Adaptation: Content-based adaptation at the event level based on MPEG2 has been implemented using the codec of the *MPEG Software Simulation Group* open source library vers. 12 (http://www.mpeg.org/MPEG/MSSG), referred to in the following as S-MPEG2. The S-MPEG2 codec performs frame resizing with bilinear interpolation and selective event compression by applying the same quantization scale to all the macroblocks of all the frames of the highlight (the quantization scale is coded with 5 bits). With the test set, we have measured that S-MPEG2 requires approximately the average bandwidth allocation of 80 Kbits, with average bandwidth allocations of 40.9, 76.2, and 185 Kbits, respectively, for the video clips in the three classes. For the same test set, with standard MPEG2, the same average bandwidth of 80 Kbits is obtained at a quantization scale of about 18, applied to all the clips, irrespectively of their content. The improvement in PSNR that is obtained with S-MPEG2 wrt standard coding is about 9% for the events in the highest relevance class.

S-MPEG2 supports also the application of different coding at the object level, by applying a different quantization scale to the DCT quantization matrix of each frame macroblock. If two or more objects are present in the same macroblock, that macroblock is associated with the quantization scale of the object with the highest relevance. Fig. 2(a) provides a comparative display of the bandwidth requirements of coding at the event, object-event, and object level, with S-MPEG2, for a sample clip with two highlights. With coding at the event level, a low compression rate is applied to each frame for the duration of the two events; with coding at the object-event level, low compression is applied only to the objects (the players in this case), of the frames of the two highlights; with coding at the object level, players have always a low compression. Fig. 2(b) displays the viewing quality achieved in the three cases, as observed at sample frames. In our implementation, to obtain more effective adaptation at the object level, the blobs of the relevant objects that are detected are enlarged with a surrounding *aura* that is displayed at the same resolution of the object. Following the model of foveation of the human vision system, assuming that the display is observed from a distance of about 40 cm [30] and taking the Sharp Zaurus SL-C700, as the reference device, the aura applied to each boundary pixel of the object has been defined equal to 30 pixels. From images labeled 1 and 2 it can be noticed that, with S-MPEG2, there is no appreciable difference in viewing quality between coding at the event and object-event level, although there is a significant difference in bandwidth allocation.

2) MPEG-4 Based Adaptation: Content-based adaptation based on MPEG4 generally achieves better results than MPEG2. Adaptation at the event level has been implemented with a modified version of the Xvid open source software (http://www.xvid.org) of the MPEG4 Simple Profile (referred to in the following as S-MPEG4-SP). Similarly to S-MPEG2, S-MPEG4-SP permits different quantization scales in different frames. It also includes texture coding, AC coefficient prediction and advanced motion vector prediction. With these improvements, the request of bandwidth is reduced with respect to S-MPEG2. With the test set, S-MPEG4-SP requires approximately the average bandwidth allocation of 21.5 Kbits, with

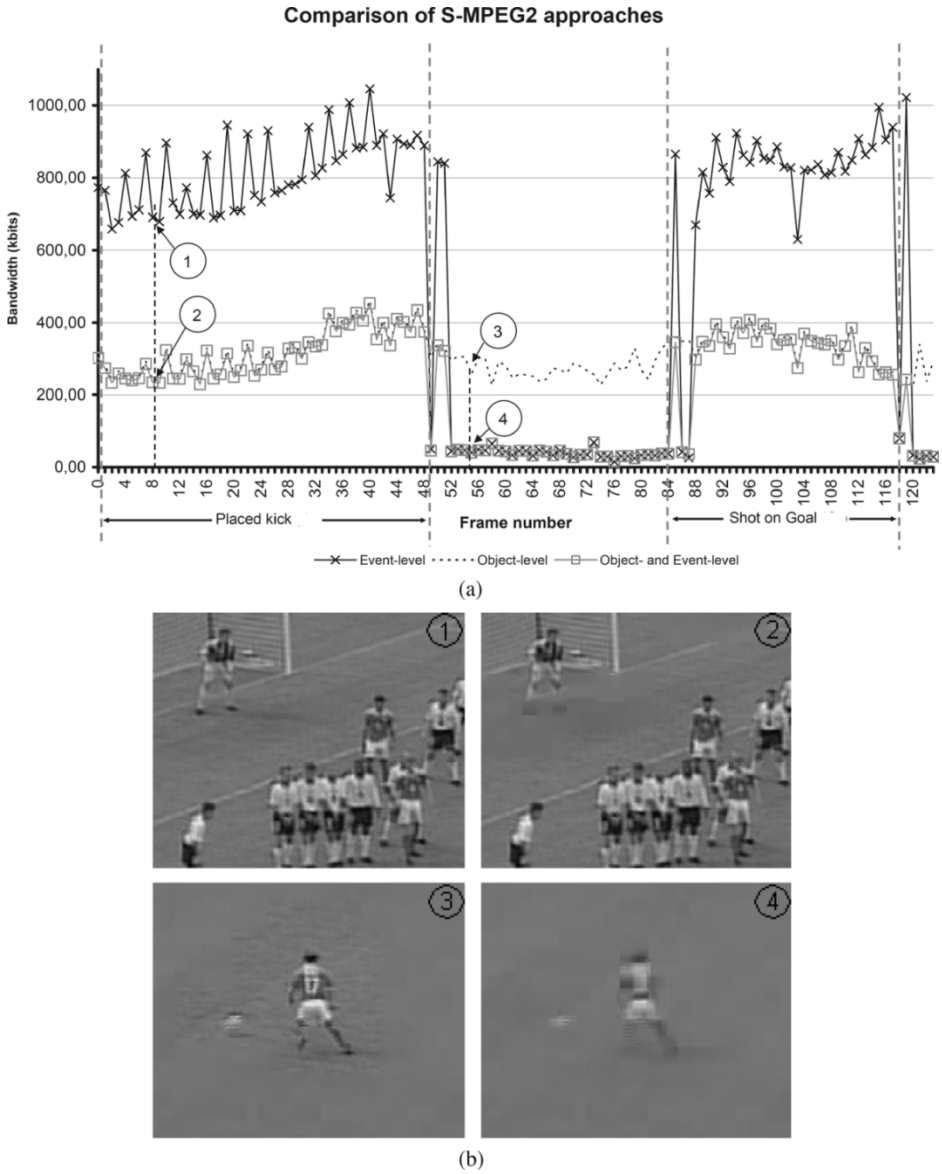


Fig. 2. S-MPEG2 adaptation at the event, object, and object-event level. a) bandwidth allocation at fixed viewing quality; b) sample frames to compare the quality achieved with the three compression policies: event (1) and object-event level (2), object (3) and object-event level (4).

average bandwidth allocations of 11.2, 29.3, and 67.4 Kbits, respectively, for the events in the three classes. For the same test set, with standard MPEG4, the same average bandwidth of 21.5 Kbits is obtained at a quantization scale of about 16, applied to all the clips, irrespectively of their content. Similar PSNRs are obtained for S-MPEG4-SP and S-MPEG2 compression at event level, at very different bandwidths: S-MPEG4-SP requires on average about one third of S-MPEG2 bandwidth. Like with S-MPEG2, improvement in PSNR with semantic adaptation with S-MPEG4-SP is about 9% wrt standard coding.

Since S-MPEG4-SP does not allow different quantization scales within the same frame, it cannot be used for object and object-event level adaptation. According to this, content-based compression for MPEG4 at object-event level has been implemented considering the MPEG4 core profile codec (referred to in the following as S-MPEG4-CP) based on the Xvid open source software [31]. S-MPEG4-CP supports adaptation at the object level of objects of any arbitrary shape. The original video stream is divided into secondary streams, each of which is associated to a distinct object. Each secondary stream is handled by a different encoder that encodes the temporal evolution of the visual object, disregarding its relationships with the other objects and the background. Different quantization scales can be therefore applied to distinct objects, depending on their relevance. The encoded streams are finally multiplexed in a single stream. However, with MPEG4, we have verified that there is almost no difference in bandwidth allocation between S-MPEG4-CP based compression at the object-event level and S-MPEG4-SP based compression at the event level, for highlight clips with fast camera motion and large objects (for example the playfield). This is due to the fact that S-MPEG4-CP has been originally conceived for object animation and graphical manipulation, and has therefore poor performance in the presence of large objects and fast camera motion. In fact, in this case, the object shape changes from one frame to the following, alpha planes are therefore large and determine some bandwidth

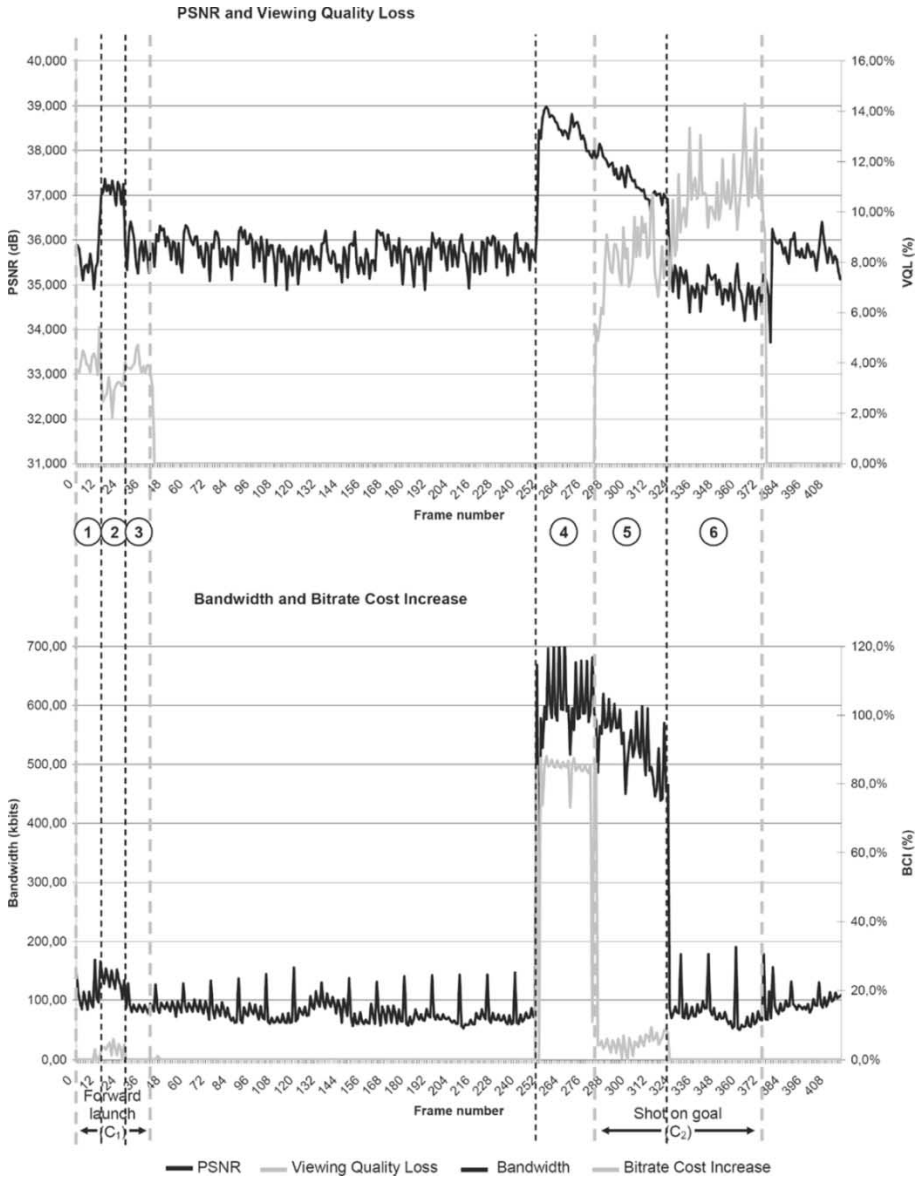


Fig. 3. Comparison between PSNR and BR classical metrics and newly defined VQL and BCI.

waste, and there is no possibility of predicting the future aspect of the object shapes.

III. Performance Measures for Semantic Adaptation

Objects and events that have the same degree of interest can be aggregated into a finite number N_{cl} of *Classes of Relevance*. Each class of relevance C is defined as

$$C = \langle \mathbf{e}, \mathbf{o} \rangle \quad with \quad \mathbf{e} \subseteq \mathbf{E}, \mathbf{o} \subseteq \mathbf{O} \tag{1}$$

where $\mathbf{E}$ and $\mathbf{O}$ are respectively the set of event types e_i and of object types o_i that the annotation engine is capable to detect, and $\mathbf{e}$ and $\mathbf{o}$ are respectively subsets of $\mathbf{E}$ and $\mathbf{O}$, such that their elements have the same degree of interest for the user. According to this definition, for the duration of the events of type $e_i \in \mathbf{e}$, only objects of type $o_j \in \mathbf{o}$ are compressed at the quantization scale decided for class C. The pair $\langle \mathbf{e}, \mathbf{o} \rangle$ of class C cannot be assigned to another class. The pair $\langle \mathbf{E}, \mathbf{o} \rangle$ indicates that all the objects of any type $o_j \in \mathbf{o}$ have the same compression ratio for any event, while, correspondingly, $\langle \mathbf{e}, \mathbf{O} \rangle$ indicates that the same compression ratio applies to the whole frame for the entire duration of the events of any type $e_j \in \mathbf{e}$. One special class, the *residual* class C_0 is associated to frames that contain events and objects not of user's interest or that cannot be detected by the annotation. Weights of relevance $w_0, w_1, \ldots, w_{N_{cl}}$, with $w_i \in [0,1]$, can be defined so that adaptation is performed with a compression ratio proportional to the w_i weight, and the smallest compression according to the bandwidth available is applied to the class of highest relevance. Thus, for instance, if relevance weights (0.1,0.5,1) are applied to classes (C_0, C_1, C_2) the quality of the elements of class C_0 after compression is expected to be approximately ten times lower than the quality of elements of class C_2.

Due to the presence of several classes of relevance, errors in the classification of objects and events made in the annotation eventually reflect into a compression ratio different than expected. 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) are produced at a higher viewing quality, and will have transmission costs higher than expected;

More precisely, for each frame I^t of the sequence produced by the annotation, we can distinguish different sets of pixels:

- $NoErr^t$: it is the set of pixels that have been correctly classified (either the whole frame in the case of correct event classification or the pixels of the objects that have been correctly classified, when objects are considered);
- $Err^t_Q = \{p \in I^t : (p \in Err^t_{Q_e}) \vee (p \in Err^t_{Q_o})\}$: it is the set of under-estimated pixels, where $Err^t_{Q_e} \Leftrightarrow (E_u \vee E_m)$ is the whole frame I^t in the case of under-estimated or missed event, and $Err^t_{Q_o} \Leftrightarrow (E_c \wedge (O_u \vee O_m))$ is the set of pixels of the objects that are either under-estimated or missed for a correctly classified event, when objects are considered; both these cases determine loss of viewing quality;
- $Err^t_C = \{p \in I^t : (p \in Err^t_{C_e}) \vee (p \in Err^t_{C_o})\}$: it is the set of over-estimated pixels, where $Err^t_{C_e} \Leftrightarrow (E_o \vee E_f)$ is the whole frame I^t for the events that are either over-estimated or falsely detected; whereas $Err^t_{C_o} \Leftrightarrow (E_c \wedge (O_o \vee O_f))$ is the set of pixels of objects that are over-estimated or falsely detected in frames associated with events correctly classified, when objects are considered; both these cases determine bandwidth waste and therefore increase of transmission costs.

Using the definitions above and the definitions of PSNR and BR, we can derive the following new indices of performance that measure the effects of annotation over adaptation, in relationship with user's preferences:

- *Viewing Quality Loss* (VQL): Resulting from over-compression due to under-estimation and miss conditions occurred in the annotation;
- *Bit-rate Cost Increase* (BCI): Resulting 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 and the PSNR measured over the same set of pixels in the ideal case of no annotation errors:

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

The PSNR of the set Err^t_Q in the case of nonnull annotation errors is obtained as

$$PSNR_{Err^t_Q} = 10 \log_{10} \left(\frac{V^2_{MAX}}{MSE_{Err^t_Q}} \right) \tag{3}$$

where V_{MAX} is the maximum (peak-to-peak) value of the signal to be measured and $MSE_{Err^t_Q}$ is the Mean Square Error calculated for the set of pixels Err^t_Q, defined as follows:

$$MSE_{Err^t_Q} = \frac{\sum_{p \in Err^t_Q} d^2(p)}{\left| Err^t_Q \right|} \tag{4}$$

being $d(p)$ the *Euclidean distance* in the RGB color space, that measures the error between the original and the distorted image. In the ideal case of error free annotation, $PSNR^{ID}_{Err^t_Q}$ only records the degradation in quality due to the compression standard and the quantization scale adopted. Since $PSNR_{Err^t_Q}$ is lower or equal to $PSNR^{ID}_{Err^t_Q}$, their ratio in (2) is always between 1 (ideal annotation case) and 0 (maximum distortion case, due to the annotation and adaptation processes).

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:

$$BCI^t = 1 - \frac{BR^{ID}_{Err^t_C}}{BR_{Err^t_C}}. \tag{5}$$

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:

$$VQL = \frac{\sum_{t=0}^{N} VQL^t}{N}; \quad BCI = \frac{\sum_{t=0}^{N'} BCI^t}{N'} \tag{6}$$

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.

Fig. 3 shows PSNR, BR, VQL and BCI for a sample clip. From this figure, the additional information conveyed by VQL and BCI, wrt PSNR and BR, in the presence of semantic adaptation, is clearly visible. In this example, adaptation is performed at the event and object level. The two events (one Forward launch between frames 0 and 42, and one Shot on goal between frames 282 and 375) have been associated to two distinct classes of relevance, together with the Playfield object. Other highlights and nonplayfield frame pixels have been associated with the *residual* class. As it can be noticed from the figure, the annotation engine detects a Forward launch in the frame interval 12–26 (thus with event frames missed in the intervals marked with labels 1 and 3) and a Shot on goal in the frame interval 251–322 (thus with event frames falsely detected and missed in the intervals labeled 4 and 6, respectively). There are also some errors in Playfield segmentation between frames 24 and 36, and 288 and 324.

Plottings of VQL and BCI allow to distinguish the effects of misses from those due to falses and underestimation of the playfield size and at the same time to view the effects of having different relevance weights. In particular, effects of Playfield underestimation are clearly evidenced in the BCI plotting (the small values of BCI recorded in intervals 2 and 5). The corresponding increase of VQL (especially visible in interval 5) provides a measure of the impact of these errors on the visual appearance of the clip. Instead, from PSNR and BR only, it is not possible to understand how much of the PSNR decrease is due to annotation errors and how much is due to playfield underestimation because of segmentation errors. The VQL observed in

TABLE VI
USER PROFILES USED TO EVALUATE THE IMPACT OVER VQL AND BCI OF CHANGES OF USER'S PREFERENCES. HIGHLIGHTS ARE INDICATED WITH THEIR INITIALS

USER PROFILE	$CLASS\ C_2$	$CLASS\ C_1$	$CLASS\ C_0$	RELEVANCE WEIGHTS
Reference	< { SG,FL} , * >	< { PK,AA} , players >	residuals	(1.0, 0.3, 0.1)
A	< SG, * >	< { FL, PK, AA} , players >	residuals	(1.0, 0.3, 0.1)
B	< { SG,FL} , * >	< { PK, AA} , players >	residuals	(1.0, 0.6, 0.5)
C	< { SG, FL} , players >	< { PK, AA} , players >	residuals	(1.0, 0.3, 0.1)
D	< { SG, FL} , { playfield, players} >	< { PK, AA} , players >	residuals	(1.0, 0.3, 0.1)

intervals 1, 3, and 6 provides a measure of the visual impact of the annotation misses. The high value of BCI in interval 4 is instead due only to false detections in this interval.

From the definitions, general criteria are immediately derived to ensure semantic adaptation with small VQL and BCI. In particular, in order to have minimum VQL, highly frequent events with high miss rate should be clustered into a class of relevance with a relevance weight close to that assigned to the *residual* class. In fact, since the frames of a missed event are compressed at the rate of the *residual* class, the VQL is small. For the same reason, frequent events with high mutual misclassification rate should be assigned to the same class, or to classes with close relevance weights, or to distinct classes of relevance, with the event of highest misclassification rate assigned to the class of lower relevance. Analogously, minimum BCI is obtained when highly frequent events with high false detection rate are assigned to a class of relevance with a relevance weight close to that assigned to the *residual* class. Very frequent events with high mutual misclassification rates should be assigned to the same class, or to distinct classes with similar relevance weights, or to distinct classes of relevance, with the event of highest misclassification rate assigned to the class of higher relevance. Objects with high false detection rate should be assigned to a class with a relevance weight close to that of the *residual* class. The increase of VQL and BCI can be predicted, for user preferences with highlight/object class assignments that do not follow these indications.

IV. EXPERIMENTS AND PERFORMANCE ANALYSIS

In the following, we present experiments of semantic adaptation of soccer video and provide some performance figures. We allowed to interactively define personalized user profiles, by changing the assignments of highlights and objects to the classes of relevance. We have assumed that users are mainly interested in those highlights that might lead to score a goal, namely, Shot on goal, Attack action, Placed kick, and Forward launch. Therefore only these events, among those that are detected by the annotation engine, have been cited in the user profiles.

To show the effects of the performance of the annotation engine in relationship with the frequency/duration of highlights and their assigned relevance, we have defined several distinct user profiles as listed in Table VI. The reference profile corresponds to the typical ranking of relevance for highlights and objects, in soccer. Clustering of highlights and objects into the classes of relevance has been made following the general principles indicated in the previous Section, taking into account misclassification and miss rates of the annotation engine (Tables I–III and V) and the average frequency of occurrence of each highlight, as from Table IV. The other profiles correspond to rankings of relevance among the most frequently defined by users.

In more detail, Profile A is used to show the effect of moving a frequent event (Forward launch) with low misclassification and miss rate into a class of lower relevance (specifically, player pixels of Forward launch frames are moved from class C_2 of weight 1, into class C_1 of weight 0.3; remaining pixels of Forward launch frames are moved into class C_0 of weight 0.1). Profile B has been defined in order to show the effect of reducing the quantization scale between the classes of relevance (from 1, 0.3, 0.1 to 1, 0.6, 0.5); Profile C permits to observe the difference in performance that occurs when compression at the class of highest relevance is applied only to objects whose size is small with respect to the frame size (only to players of the highlights in class C_2, instead of the entire frame). Profile D is used to show the effect when compression is applied to large objects at the highest relevance class (to both the players and the playfield of the highlights in C_2).

Performance has been analyzed with reference to a test set appropriately built so as to derive figures that are close to those observable by users, when viewing a full soccer game in reality. In particular, we have collected 35 Forward launch, 20 Shot on goal, 12 Attack action, and 15 Placed kick video clips (in total 8876 frames of standard PAL 720×576). From this initial set, to have performance figures that are statistically meaningful, we have created ten distinct test subsets, each of which includes 25 Forward launch, 14 Shot on goal, eight Attack action, and 11 Placed kick clips, randomly extracted from the clips of the initial set. The highlights in the test subsets have the same relative frequency of occurrence as in a soccer game in reality (see UEFA statistics in Table IV). Since the annotation engine performs errors in both the classification of highlights, and in the detection of the frames of the highlights correctly detected, for each test subset and for each highlight, we have added a number of clips so as to obtain missed and false events in the same proportions as those measured for the annotation engine, indicated in Tables I and II. For each user profile, the average values of PSNR, BR, VQL, and BCI have been finally obtained by averaging over the ten subsets the corresponding measures of each subset. Semantic adaptation has been performed at object-event level, using the S-MPEG2 codec. We have considered as relevant objects playfield and players. Video data have been transcoded with standard downscaling in order to adapt them to the Sharp Zaurus SL-C700 display. Table VII reports the average values of BR, PSNR, VQL, and BCI that have been measured.

It can be noticed that PSNR and BR mirror the requirements of the different user profiles. PSNR and BR of profile A are lower than those of the reference profile, due to the fact that

TABLE VII
BIT RATE, PSNR, VQL, AND BCI AVERAGED OVER THE TEN TEST SETS

	AVERAGE PERFORMANCE MEASURES				
	Reference profile	Profile A	Profile B	Profile C	Profile D
BR (Kbps)	3742.33	1870.07	3839.55	1028.64	2664,06
PSNR (dB)	33.51	31.97	34.13	31.35	32.20
VQL	2.23%	0.84%	1.48%	1.36%	2.56%
BCI	6.11%	2.65%	5.73%	3.49%	6.31%

only Shot on goal events are associated with the highest relevance class. The saving in bandwidth is sensible. On the other hand, PSNR and BR of profile B are higher due to the fact that in this profile, lower classes have higher relevance weights. In profiles C and D, for each class of relevance, compression is only applied to objects instead of the entire frames. It can be noticed that this reflects into the fact that their BR and PSNR have lower values than the reference profile. In particular, BR is significantly reduced (by 72% considering players only, and 28% when both players and playfield are considered). PSNR has a smaller reduction (about by 6% for players and 4% for players and playfield), mainly because S-MPEG2 applies compression to macroblocks instead of object pixels. The higher BR and PSNR in profile D wrt C is due to the fact that in D objects encoded in high quality are larger.

Low values of VQL and BCI observed for profiles A and B are essentially due to the fact that, according to Table I, event misclassifications have almost no influence: events of class C_2 are never classified into events of class C_1 and vice versa; only Attack action clips in class C_1 have relatively high probability to be classified into clips of class C_0, but this misclassification determines a small compression difference. Very low VQL and BCI of profile A are the consequence of moving Forward launch into a class of lower relevance: results indicate that an average reduction of VQL and BCI values by 62% and 56%, respectively, can be expected. Reduction of the quantization scale between the classes, as in profile B, also reduces sensibly the impact of annotation errors and particularly over VQL (an average decrease of VQL by 33% is estimated); in fact the effects of misses are weighted less. Finally, profile C shows, as expected, that errors in the segmentation of small objects (players) have much lower impact than errors in event classification: VQL and BCI are reduced by 39% and 42%, respectively. Instead, profile D indicates that errors in the segmentation of large objects, in that they exist also in all the frames of the events correctly detected, have great impact: VQL and BCI of profile D are the highest ones measured. Overall, profile A and C provide a good trade-off between bandwidth allocation requirement, viewing quality and minimization of the effects of the errors in the annotation. In Fig. 4, we report a short sequence of adapted video for the Sharp Zaurus SL-C700 as it appears in the presence of event underestimation with estimated viewing quality loss of 2.56%. We can observe that this percentage of viewing quality loss provides a strong negative perceptual impact.

BR, PSNR, VQL, and BCI have also been computed for each highlight, so as to understand how characteristics and statistics of the highlights influence the performance of semantic adaptation, in average. Table VIII presents the average values of BR and PSNR for each profile and for each highlight, as measured

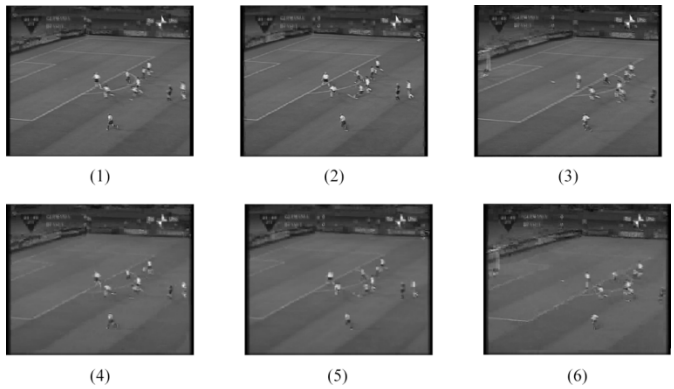


Fig. 4. Perceptual effect of event underestimation with 2.56% viewing quality loss for a sample clip.

over the whole test set, while the effects of erroneous classification on VQL and BCI can be observed in Table IX. From the reference profile, we can notice that at the same quantization scale, Shot on goal has slightly higher BR than Forward launch (about by 2%), due to the fact that, typically, in Shot on goal, motion vectors change sensibly from one frame to the other. PSNR is lower (about by 5%), due to the fact that the variability of motion vectors determines errors in their prediction. This is always true except for profile A, where Forward launches are associated with a class of lower relevance so that less viewing quality loss is determined by misses.

The average VQL and BCI of Shot on goal are respectively lower (about by 31%) and larger (about by 51%) than VQL and BCI of Forward launch. Lower VQL is explained with the fact that, although Shot on goal has higher miss rate and lower false detection rate than Forward launch (see Tables I and II), it has much less misses at the frame level (see Table III), so that the average total number of missed frames per event is sensibly lower. Higher BCI is due to the fact that, although the number of falses is almost the same as Forward launch, Shot on goal has higher motion and therefore annotation errors reflect into higher cost of the adaptation. The effects of the errors of object segmentation are visible by analysing the VQL and BCI values of Forward launch and Shot on goal of profile C wrt profile D: errors in the segmentation of large objects (i.e., playfield) result into larger error values.

From profile A we can observe that Forward launch has higher BR than Attack action and Placed kick (respectively by 26% and 63%). Attack action has always higher BR than Placed kick (about by 29%). The difference of BR between Attack action and Placed kicks can be explained by the different behavior of adaptation in the two cases, due to the fact that the first event is fast paced and has usually larger camera displacements, while the second is instead almost motionless. PSNR values are instead similar (approximately a 4.5% difference) due to the high compression rate applied. In profile B this percentage lowers to 16.7% for BR and 4% for PSNR. Attack action has always higher VQL than Placed kick (about 5 to 1 ratio), due to the higher percentage of event misses that are observed (see Table II), and the higher number of frame misses. BCI are instead very similar, due to the fact that although Attack action has a total number of falsely detected frames (due to false event classification and false frame detection within the event) lower

TABLE VIII
BIT-RATE AND PSNR (VALUES ARE IN kbps FOR BR AND dB FOR PSNR). HIGHLIGHTS ARE INDICATED WITH THEIR INITIALS

	PSNR					BIT-RATE				
	Reference	Profile A	Profile B	Profile C	Profile D	Reference	Profile A	Profile B	Profile C	Profile D
FL	35.23	31.67	35.30	31.79	33.04	5091.84	773.82	5125.12	1198.45	3628.22
SG	33.44	33.44	33.71	31.00	32.14	5205.69	5205.69	5245.86	1408.56	3536.11
AA	31.73	31.73	33.46	31.73	31.73	610.35	610.35	835.75	610.35	610.35
PK	30.35	30.35	32.13	30.35	30.35	473.65	473.65	717.56	473.65	473.65

TABLE IX
VIEWING QUALITY LOSS AND BIT-RATE COST INCREASE. HIGHLIGHTS ARE INDICATED WITH THEIR INITIALS

	VQL					BCI				
	Reference	Profile A	Profile B	Profile C	Profile D	Reference	Profile A	Profile B	Profile C	Profile D
FL	4.03%	0.54%	2.48%	2.13%	4.46%	7.17%	0.42%	6.56%	3.14%	7.00%
SG	2.87%	2.87%	1.79%	2.05%	3.15%	10.82%	10.82%	10.08%	7.77%	11.40%
AA	1.77%	1.77%	1.48%	1.77%	1.77%	0.40%	0.40%	0.70%	0.40%	0.40%
PK	0.34%	0.34%	0.34%	0.34%	0.34%	0.37%	0.37%	0.47%	0.37%	0.37%

than that of Placed kick, indeed they require a higher BR per frame so that errors result into higher BCI. The only exceptions are observed in profile B where VQL is lowered, as a side effect of the fact that missed and under-estimated events and frames are weighted less (the difference between relevance weights is 0.1, instead of 0.2), and the same happens to the effect of misclassifications into events of a lower class. BCI instead is increased due to the fact that higher bandwidth is associated with missed and under-estimated frames (compression is proportional to 0.5 instead of 0.1). This effect overcomes the savings in bandwidth that result from the few false detections at the event and frame level.

From the results obtained, we can observe that Shot on goal and Forward launch are the most critical highlights for semantic adaptation of soccer video: in that they model complex actions, they require high BR per frame; moreover they have respectively high BCI (due to high false rate) and high VQL (due to high miss rate). In order to have them in the highest class of relevance, their annotation error rates must be reduced, so as to decrease their impact on BCI and VQL. Improvement of Attack action miss rate is also needed if Attack action is assigned to classes of high relevance. The sole improvement of object segmentation has no substantial impact on the performance of semantic adaptation.

V. CONCLUSIONS

In this paper, we have discussed semantic adaptation from the view point of the impact that the errors performed by the annotation engine have over the system performance. From PSNR and Bit Rate, we have defined two new performance measures for semantic adaptation, namely Viewing Quality Loss and Bit-rate Cost Increase, that can be used to measure the viewing quality loss and the bandwidth waste in reference to user's preferences and expectations, and are in direct relationship respectively with under-estimation and missing and over-estimation and false detection that occur in the annotation.

General guidelines for the design of user preferences can be derived. Once the most interesting highlights are selected among those that can be detected by the annotation engine, both the performance figures of the annotation engine and the frequency/duration of highlights must be considered for the definition of user preferences. Low VQL and BCI are achieved if frequent events with respectively high miss rate and high false detection rate are associated with a class of relevance with a relevance weight close to that assigned to the residual class. Quantitative indications on how much annotation errors reflect into the output of semantic annotation have been given with reference to a prototype system for semantic adaptation of soccer video.

REFERENCES

[1] R. Mohan, J. Smith, and C. Li, "Adapting multimedia internet content for universal access," *IEEE Tran. Multimedia*, vol. 1, no. 1, pp. 104–114, Mar. 1999.
[2] B. L. Tseng, C.-Y. Lin, and J. R. Smith, "Using MPEG-7 and MPEG-21 for personalizing video," *IEEE Multimedia*, vol. 11, no. 1, pp. 42–52, Jan.–Mar. 2004.
[3] R. Cucchiara, C. Grana, and A. Prati, "Semantic video transcoding using classes of relevance," *Int. J. Image Graph.*, vol. 3, no. 1, pp. 145–169, Jan. 2003.
[4] A. Vetro, "MPEG-21 digital item adaptation: enabling universal multimedia access," *IEEE Multimedia*, vol. 11, no. 1, pp. 84–87, Jan.–Mar. 2004.
[5] Y.-F. Ma, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in *Proc. ACM Int. Conf. Multimedia*, 2002, pp. 533–542.
[6] O. Werner, "Requantization for transcoding of MPEG-2 bit streams," *IEEE Trans. Image Process.*, vol. 8, no. 2, pp. 179–191, Feb. 1999.
[7] G. Keesman, R. Hellinghuizen, F. Hoeksema, and G. Heideman, "Transcoding of MPEG bitstreams," *Signal Process.: Image Commun.*, vol. 8, no. 6, pp. 481–500, Sept. 1996.
[8] T. Shanableh and M. Ghanbari, "Heterogeneous video transcoding to lower spatio-temporal resolution and different encoding formats," *IEEE Trans. Multimedia*, vol. 2, no. 2, pp. 101–110, Jun. 2000.
[9] Z. Lei and N. Georganas, "H.263 video transcoding fo spatial resolution downscaling," in *Proc. Int. Conf. Information Technology: Coding and Computing*, 2002, pp. 425–430.
[10] Y. Liang and Y.-P. Tan, "A new content-based hybrid video transcoding method," in *Proc. IEEE ICIP*, 2001, vol. 1, pp. 429–432.
[11] C.-Y. Lin, B. Tseng, and J. Smith, "Universal MPEG content access using-compressed-domain system stream editing techniques," in *Proc. IEEE Int. Conf. Multimedia & Expo*, 2002, vol. 2, pp. 73–76.
[12] A. Vetro, T. Haga, K. Sumi, and H. Sun, "Object-based coding for long-term archive of surveillance video," in *Proc. IEEE Int. Conf. Multimedia & Expo*, 2003, vol. 2, pp. 417–420.
[13] C.-Y. Lin, B. Tseng, M. Naphade, A. Natsev, and J. Smith, "Videoal- a novel end-to-end MPEG-7 video automatic labeling system," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2003, vol. 2, pp. 53–56.
[14] K. Nagao, Y. Shirai, and K. Squire, "Semantic annotation and transcoding: making web content more accessible," *IEEE Multimedia*, vol. 8, no. 2, pp. 69–81, Apr.–June 2001.
[15] C. Shih-Fu, D. Zhong, and R. Kumar, Real-Time Content-Based Adaptive Streaming of Sports Video Columbia Univ., New York, ADVENT Tech. Rep. 121, Jul. 2001.
[16] T. Ebrahimi and M. Kunt, "Visual data compression for multimedia applications," *Proc. IEEE*, vol. 86, no. 6, pp. 1109–1125, Jun. 1998.

[17] A. Cavallaro, O. Steiger, and T. Ebrahimi, "Semantic segmentation and description for video transcoding," in *Int. Conf. Multimedia & Expo (ICME), Special Session on Video Segmentation for Semantic Annotation and Transcoding*, 2003, pp. 597–600.
[18] M. Kunt, "Object-based video coding," in *Handbook of Image and Video Processing*. New York: Academic, 2000, ch. 6.3, pp. 585–596.
[19] 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 Trans. Image Process.*, vol. 9, no. 4, pp. 636–650, Apr. 2000.
[20] Z. Wang, A. C. Bovik, and L. Lu, "Why is the image assessment so difficult?," in *Proc. IEEE Conf. Acoustics, Speech, and Signal Processing*, May 2002.
[21] M. Bertini, R. Cucchiara, A. D. Bimbo, and A. Prati, "An integrated framework for semantic annotation and transcoding," *Multimedia Tools Applicat.*, to appear.
[22] J.-G. Kim, Y. Wang, and S.-F. Chang, "Content-adaptive utility-based video adaptation," in *Proc. IEEE Int. Conf. Multimedia & Expo*, Jul. 2003, pp. 281–284.
[23] G. Baldi, C. Colombo, and A. Del Bimbo, "A compact and retrieval-oriented video representation using mosaics," in *Proc. 3rd Int. Conf. Visual Information Systems (VISual99)*, Jun. 1999, pp. 171–178.
[24] R. Nelson, "Finding line segments by stick growing," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 5, no. 16, pp. 519–523, May 1994.
[25] S. Intille and A. Bobick, "Recognizing planned, multi-person action," *Comput. Vis. Image Understand.*, vol. 81, no. 3, pp. 414–445, March 2001.
[26] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[27] J. Assfalg, M. Bertini, C. Colombo, A. D. Bimbo, and W. Nunziati, "Semantic annotation of soccer videos; automatic highlights identification," *Comput. Vis. Image Understand.*, vol. 92, no. 2–3, pp. 285–305, Nov.–Dec. 2003.
[28] [Online]. Available: http:\\www.uefa.com
[29] R. Cucchiara, C. Grana, and A. Prati, "Semantic transcoding of videos by using adaptive quantization," *J. Internet Technol., Special Issue on Real Time Media Delivery Over the Internet*, vol. 5, no. 4, pp. 31–39, 2004.
[30] 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 SAE International, Tech. Rep. 2000-01-2202, Jun. 2000.
[31] M. Bertini, A. Del Bimbo, R. Cucchiara, and A. Prati, "Object-based and event-based semantic video adaptation," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2004.

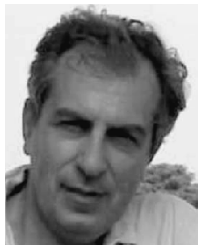
Marco Bertini received the M.S. degree in electronic engineering in 1999, and the Ph.D. degree in computer engineering in 2004, both from the University of Florence, Italy.

He carries out research activity at the Department of Systems and Informatics, University of Florence, Italy. His main research interest is content-based indexing and retrieval of videos. He is the author of more than 30 papers in international conference proceedings and journals, and is a reviewer for international journals on multimedia and pattern recognition.

Rita Cucchiara (M'92) received the Laurea degree (magna cum laude) in electronic enginering in 1989 and the Ph.D. degree in computer engineering in 1993, both from the University of Bologna, Italy.

She is currently Full Professor of computer engineering at the University of Modena and Reggio Emilia, Modena, Italy. She was formerly Assistant Professor (1993–1998) at the University of Ferrara, Italy, and Associate Professor (1998–2004) at the University of Modena and Reggio Emilia. She is currently with the Faculty staff of computer engenering, in charge of courses on computer architectures and computer vision. Her current interests include pattern recognition, video analysis and computer vision for video surveillance, domotics, medical imaging, and computer architecture for managing image and multimedia data. She is the author and co-author of more than 100 papers in international journals and conference proceedings. She currently serves as reviewer for many international journals in computer vision and computer architecture (e.g., IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON MEDICAL IMAGING, *Image and Vision Computing*, *Journal of System Architecture*, and *IEEE Concurrency*). She is in the editorial board of *Multimedia Tools and Applications*.

Prof. Cucchiara is member of GIRPR (Italian chapter of the International Association of Pattern Recognition), AixIA (Italian Association of Artificial Intelligence), the ACM, and the IEEE Computer Society. She participated at scientific committees of the outstanding international conferences in computer vision and multimedia (CVPR, ICME, ICPR) and symposia and organized special tracks in computer architecture for vision and image processing for traffic control.



Alberto Del Bimbo (M'90) is Full Professor of computer engineering at the University of Florence, Italy, where, since 1998, he is the Director of the Master in Multimedia. At the present time, he is Deputy Rector of the University of Florence, in charge of research and innovation transfer. His scientific interests are pattern recognition, image databases, multimedia, and human–computer interaction. He is the author of over 170 publications in the most distinguished international journals and conference proceedings. He is the author of the *Visual Information Retrieval* monography on content-based retrieval from image and video databases (Morgan Kaufman).

Prof. Del Bimbo is a Fellow of the International Association for Pattern Recognition (IAPR). He is presently Associate Editor of *Pattern Recognition, Journal of Visual Languages and Computing, Multimedia Tools and Applications Journal, Pattern Analysis and Applications*, IEEE TRANSACTIONS ON MULTIMEDIA, and IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He was the guest editor of several special issues on image databases in highly respected journals.

Andrea Prati (M'98) received the Laurea degree (magna cum laude) in computer engineering in 1998 and the Ph.D. degree in computer engineering, both from the University of Modena and Reggio Emilia, Italy, in 2002). During the final year of his Ph.D. studies, he spent six months as visiting scholar at the Computer Vision and Robotics Research (CVRR) Lab at the University of California, San Diego (UCSD), working on a research project for traffic monitoring and management through computer vision.

He is currently an Assistant Professor with the Faculty of Engineering, Dipartimento di Scienze e Metodi dell'Ingegneria, University of Modena and Reggio Emilia. His research interests are mainly on motion detection and analysis, shadow removal techniques, video transcoding and analysis, computer architecture for multimedia and high performance video servers, video-surveillance and domotics. He is author of more than 60 papers in international and national conference proceedings and leading journals and he serves as reviewer for many international journals in computer vision and computer architecture.

Dr. Prati is a member of the ACM and IAPR.