Semantic Video Adaptation based on Automatic Annotation of Sport Videos

Marco Bertini, Alberto Del Bimbo
Dipartimento di Sistemi e Informatica
University of Florence
Via S. Marta, 3
Firenze, Italy
{bertini,delbimbo}@dsi.unifi.it

Rita Cucchiara, Andrea Prati
Dipartimento di Ingegneria dell'Informazione
University of Modena and Reggio Emilia
Via Vignolese, 905
Modena, Italy
{cucchiara.rita, prati.andrea}@unimore.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 precision of the automatic annotation and the way of operation of the codec which is used to perform adaptation at the event or object level. In this paper, we discuss critical factors that affect performance of automatic annotation and define new performance measures of semantic adaptation, *Viewing Quality Loss* and *Bitrate Cost Increase*, that are obtained from classical PSNR and Bit Rate, but relate the results of semantic adaptation with the user's preferences and expectations. The new measures are discussed in detail for a system of sport annotation and adaptation with reference to different user profiles.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*Systems issues*, *User issues*; H.2.4 [**Systems**]: Multimedia databases; I.4.2 [**Compression (Coding)**]

General Terms

Performance, Human factors

Keywords

Video adaptation, automatic video annotation, transcoding

1. INTRODUCTION

Universal multimedia access is becoming more and more popular due to the diffusion of new devices to access to multimedia data from any place. Among multimedia data, videos are probably the more challenging since they call for high bandwidth requirement to preserve as much as possible of the original quality. However, meeting the constraints of the device and the requirements of the user, and keeping low the costs of the transmission (in terms of data transferred and time required) at the same time, is not a trivial task.

Video adaptation techniques have been widely studied in the last years [13, 8] in order to enable Universal Multimedia Access (UMA) from any place and also with devices with limited resources. Most of the video adaptation techniques provide syntactic video adaptation performing scaling, color subsampling, temporal downscaling or changing the compression's factor [9]. This results in that the video is adapted equally. Therefore, there is, on the one side, bandwidth waste for preserving the quality of useless parts of the video, and, on the other side, excessive degradation of meaningful parts.

As a consequence, recently many researchers have concentrated their efforts in defining new "semantics-based" or "content-based" video adaptation approaches. The rationale is that the user can elicit relevant *video elements* (either *objects* or *events* of interest) and define for each of them a degree of relevance. Relevant elements should be detected automatically in the video, possibly with computer vision-based annotation modules, and the quality of their transmission should be adapted to their user-defined relevance. This selective adaptation can be done at *object-level* (connected regions in a frame) or at *event-level* (sequences of frames with common meaning). 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 objects (e.g., regions in the surrounding of the players).

Video adaptation in terms of the relevance of the objects detected in each frame has been addressed by [14] and [2] for video surveillance applications. In [14], Vetro et al. presented an object-based transcoding framework that uses dynamic programming or metadata, for the allocation of bits among the multiple objects in the scene. In [7] the advantages of representing visual data and thus semantics in terms of regions corresponding to objects is clearly evidenced. Chang et al. [6] have filtered live video content according to events and highlights. In [2] we have developed a prototype system for annotation and adaptation of soccer sport videos, with adaptation based on objects and events. However, a still open problem is the choice of the granularity of the elements to be exploited for the adaptation, that is deciding to work at object- or event-level. A detailed comparison of the possible approaches has been discussed in [3].

In addition, there is the need of a reliable and consistent performance evaluation of content-based video adaptation systems. Most

MIR'04, October 15–16, 2004, New York, New York, USA.
Copyright 2004 ACM 1-58113-940-3/04/0010 ...$5.00.

of the measures for performance evaluation of video adaptation systems are, however, still based on the PSNR (Peak Signal-to-Noise Ratio) [6, 2] with some noticeable exceptions that take into account non-linear distortion effects on the human perception system [14, 5]. However, in the case of content-based video adaptation, they all can not take into account user's satisfaction and how much this is affected by errors in the video annotation system. A few approaches in this direction have been proposed recently. A weighted PSNR has been defined in [2] to include user's preferences. Chang et al. [6] have defined a function that takes into account both quality in the video transfer (by means of PSNR) and the consumed bandwidth (using bit rate, BR).

In this paper we present a new metric for performance evaluation of content-based video adaptation systems that takes into account the overall user's satisfaction by merging the effects of annotation errors and adaptation distortions. The new performance measures, *Viewing Quality Loss* and *Bitrate Cost Increase*, are obtained from classical PSNR and Bit Rate, but relate the results of semantic adaptation with the user's preferences and expectations. They can be used with any annotation system and only content-based adaptation module.

2. ANNOTATION AND SEMANTIC ADAPTATION SYSTEMS

The reference framework is a system resulting from the integration of an automatic annotation engine and a content-based adaptation module. Video annotation has been widely studied over the last few years, resulting in many research prototypes and several commercial tools. Among the possible application contexts, sports annotation is very widespread, due to its deployment in broadcasting, post-production logging, indexing, and so on [1, 10, 16]. Known context is usually structured in an *ontology*, the definition of which is beneficial not only in the annotation process, but also for information retrieval. When video annotation is associated with video access and delivery, and thus with content adaptation, the most common frameworks for knowledge representation come from MPEG-7 and MPEG-21 standards [12, 11]. In MPEG-7 the description schemes (DS) are modeled on XML schemas, easing the use of parsing tools for indexing, querying, and retrieving information. Furthermore, efforts have been made to standardize techniques and rules for modeling the users' requests and preferences. Recently, the MPEG-21 standardization committee has addressed the UMA-related problems by including a *Digital Item Adaptation* (DIA) section in the Part 7 of the standard (ISO/IEC 21000-7) in order to adapt the media content to the device's limitations [12].

2.1 Ontology

According to the MPEG-7 terminology, each frame of a video can be divided into *spatial segments*, which are sets of not necessarily connected pixels of a frame. Within them, we call the regions with associated semantics ROIs, *regions of interest*. Thus, we can define the set of meaningful objects of a video as

$$O = \{ROI_i\} \cup \{\widetilde{o}\} \quad ; \quad O = \{ROI_1, ROI_2, ..., ROI_n\} \cup \{\widetilde{o}\}$$

where $\widetilde{o}$ are the parts of a frame that do not belong to any ROI. The ROIs are segmented by means of visual descriptors able to extract and classify objects, and to perform temporal and spatial reasoning on the scene.

Then, we shall use the concept of *temporal segments* as defined by MPEG-7. A temporal segment in MPEG-7 is a set of not necessarily contiguous frames. We shall use the term *events* to define the types of temporal segments with a specific meaning. Unlike MPEG-7 which uses the word "event" to define the condition that connects objects to each other in any instant, in our case, "event" refers only to the continuous presence of a fact along the time sequence. In practice, we consider a set of events E defined as:

$$E = \{h_i\} \cup \{\widetilde{e}\} \quad ; \quad E = \{h_1, h_2, ..., h_m\} \cup \{\widetilde{e}\}$$

where each h_i can be viewed as a highlight, while the category $\widetilde{e}$ comprises all not relevant parts of the video. The ontology is thus defined in terms of objects, events, and their relationships as described by means of acyclic graphs.

2.2 User device's requirements

The device's requirements represent the constraints that the client device imposes on the access to the video content. For instance, the maximum resolution of the device's display limits the spatial dimension of the video. In this case, spatial downscaling is mandatory and has the positive knock-on effect of reducing the required bandwidth. Furthermore, current mobile devices have limited color resolution (typically, no more than 65,535 colors). Consequently, a reduction in color depth might be necessary in order to adapt to the handset's capabilities. Although these two alterations are unavoidable and bring benefits in terms of required bandwidth, they may also entail notable image degradation, especially with regard to color reduction. Tests on basic adaptation techniques have been carried out in [4]. Mobile devices normally have a limited available memory and a computational capability that sometimes circumscribes the possibility to run sophisticated codecs and browsers. Thus, the video adaptation server should supply different encoded versions of the video, for instance with MPEG-2 or MPEG-4 standards. Another requirement that must be taken into account is the maximum bandwidth available for the connection. Current telecommunication standards for mobile devices are GPRS (General Packet Radio System) and UMTS (Universal Mobile Telecommunications System), whose maximum dedicated bandwidths are about 115 kbps and 2 Mbps, respectively. Since typical bandwidth requirements for videos at PAL/NTSC frame rate are much higher, suitable and effective compression techniques must be employed. In particular, the selective adaptation of the compression based on content and user's interests can improve performance considerably.

2.3 User's interests

The user's interests can be basically defined in terms of viewing quality and service costs. Therefore the basic performance analysis parameters that can be used are PSNR and BR. In [15], a *utility function* has been defined showing relationships between different types of resources (bandwidth, display, etc.) and utilities (objective or subjective quality, user's satisfaction, etc.). Here, bitrate and PSNR are the straightforward parameters adopted for measuring costs and quality of the video output.

The quality adaptation can be improved by exploiting semantic annotation and user's interests. In particular, we may 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 (objects and events) with the same degree of relevance for the user. Formally, given the set of classes of relevance ordered by ascending relevance $C = \{C_0, ..., C_{N_{CL}}\}$, each element 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. In this paper, we employed three classes as an example, namely C_0, C_1 and C_2 of low, medium, and high quality, respectively. The user can assign a relative weight for each class, indicating the respective ratios in the relevance, that will basically map onto the compression levels. As an example, setting the weights to $\{w_{C_0}, w_{C_1}, w_{C_2}\} = \{0.1, 0.4, 1.0\}$ means that the quality of class C_2 should be ten times better than that of class C_0.

In this case the performance evaluation depends on the user's interests. Actually, the user can select his preferences according with the semantic of the video (e.g., s/he could be more interested in a shot of goal than a placed kick). The user gives the relative interest of each class w.r.t. the others and the degree of quality (and consequent cost) needed for the most interesting class. The system selects the compression level of the classes of relevance accordingly. Then, the final performance parameters, such as PSNR and BR, are in accordance with the user's satisfaction. Nevertheless, while the variation of PSNR and BR in function of the compression is almost known, the effects of annotation errors on the final performance is not a-priori estimable.

3. ANNOTATION AND ADAPTATION OF SOCCER VIDEOS

In sports videos, users are usually interested in watching certain areas of the images, such as the playfield or the zone around the goal box in soccer, or the zone near the start or arrival in a race. These regions of interest are extracted by the automatic annotation system for two purposes: the first one is to provide a selective compression at object level, preserving as much quality as possible for the objects in which the user is more interested; the second purpose is to use the objects as inputs for the classification of events. In fact, in sports certain events can happen only in given areas and under given conditions (think for instance to the shot on goal in soccer).

The objects that are detected and extracted in soccer videos are the playfield (PF), the players and the ball (PL):

$$O_{soccer} = \{PF, PL\} \cup \{\widetilde{o}\}$$

where $\widetilde{o}$ is the area outside the playfield (e.g., the crowd), which is of no interest for the detection of highlights nor to the viewer of the video.

The playfield shape is obtained by applying color analysis and binarization to the video frames. The frame bitmap is processed using K-fill, flood fill, followed by erosion and dilation. The shape of the playfield is represented as a polygon for the purpose of automatic annotation, while for the purpose of video adaptation, the polygon is used for soccer videos, and a bitmap representation is used for swimming videos. This difference is due to the fact that accurate detection of the playfield shape and polygonal approximation are obtained precisely if the color of the playfield area is uniform, and playfield lines and player "blobs" are of a small size: the soccer field is a typical example in which polygonal shapes can be extracted reliably in most frames. The portion of playfield that is framed (and hence the playfield zone where the play takes place) is identified by the aspect of the playfield shape and the playfield lines extracted from the edge image; recognition is performed using Naïve Bayes classifiers. Players and ball blobs are extracted by color differencing and represented as "binary blobs". Constraints on the side ratio of blobs' bounding boxes and area are used to discard non-player blobs. In order to provide users with a better understanding of the video content, the blobs of players and ball are enlarged in order to include a small part of the area around them.

The problem of modeling highlights can be seen as part of the problem of detecting special occurrences within the temporal sequences. In fact, a generic highlight can be regarded as a concatenation of consecutive phases of the competition. Each phase occurs typically in a distinct zone of the playfield, while transitions between phases are related to the movement of objects such as the athletes and/or the ball. In our approach, highlights are modeled using FSMs. Each highlight is described using a directed graph they model the relevant steps in the progression of the game or race, such as moving from one part of the playfield to another, accelerating or decelerating, etc.

Meaningful events that are extracted by the annotation subsystem are the most important highlights. In particular, for soccer, highlights that have been modeled are: forward launches (FL), shots on goal (SG), spot kicks as penalty kicks (PK), free kicks near the goal post, and corner kicks, as well as attacks actions (AA) and other plays that may lead to a shot on goal. In this paper, we use an ontology as follows:

$$E_{soccer} = \{FL, SG, PK, AA\} \cup \{\widetilde{e}\} \tag{2}$$

where $\widetilde{e}$ indicates that no highlight is present in the video stream being processed.

Table 1 reports the *Detection Rate* (DR) and *False Alarm Rate* (FAR) figures of playfield zones and players in terms of pixels classified as belonging to these objects.

Sports video	Object	DR	FAR
Soccer videos	Playfield	99,9%	0.16%
	Players	99,8%	5.51%

Table 1: Performance figures of object automatic detection over 90' of soccer video .

Table 2 reports the confusion matrix, showing the precision in highlight detection and the errors in highlight classification. The percentage in the "other" column indicates the false highlight detection. Finally, in Table 3 the percentages of miss detection are reported.

The adaptation module performs content-based video adaptation according to the bandwidth requirements and the weights of the classes of relevance. Different compression techniques have been implemented that performs coding at the semantic level.

The first one exploits the standard *adaptive quantization* of MPEG-2 to select the quantization scale QS_i ($QS_i \in [0, 31]$) of each macroblock i of each frame of the video. This approach is referred to as *S-MPEG2*. For each i, the dominant class of relevance and the corresponding QS_i are computed, depending on which objects and event are involved.

Other two coding policies have been implemented based on MPEG-4 and, particularly, on the Xvid open source software (`http://www.xvid.org`). Differently from MPEG-2, in MPEG-4 the quantization values for the macroblocks *within the same Video Object Plane (VOP)* are sent in a differential format: each value for a macroblock (except for the first) is coded as {-2,-1,1,2} with respect to the base value of the VOP. This allows MPEG-4 to reduce the bandwidth required for the adaptive quantization (2 bits for each quantization value w.r.t. 5 bits), but restricts the flexibility, practically preventing us from the use of different quantization scales for the macroblocks.

The most straightforward way is to employ the MPEG-4 Simple Profile (*S-MPEG4-SP*): it does not consider objects (and, thus, does not allow different quantization factors within the same frame), but only events, i.e. different quantization scales are used in different groups of frames. Instead, working at object-level, the Core Profile of MPEG-4 can be used (*S-MPEG4-CP*) and creating a different VOP for each object extracted by the annotation system. In this

	True highlight				
Recog. highlight	Fwd.Launch	Shot Goal	Placed kick	Attack act.	Other
Fwd.Launch	89.75%	1.67%	0%	0%	8.58%
Shot on goal	1.525%	93.9%	0%	0%	4.575%
Placed kick	0%	0%	89.75%	0%	10.25%
Attack action	1.6%	1.0%	0%	97.4%	1.0%

Table 2: Performance figures of highlight automatic detection over 90' of soccer video: precision and misclassification errors

	True highlight			
Recognized highlight	Fwd.Launch	Shot on goal	Placed kick	Attack act.
Misses	5%	13%	7%	25%

Table 3: Soccer highlight misses percentage

Compr. Techn.	Avg. bandwidth	Standard	Semantic
MPEG-2	530.30 kbps	32,67 dB	35,57 dB
MPEG-4	179,94 kbps	33,47 dB	36,22 dB

Table 4: Average PSNR for MPEG-2 and MPEG-4, both standard and semantic approaches over 90' of soccer video

way, we can assign different quantization scales to each object in dependence to its relevance for the user. However, this approach has proven to be not suitable in the case of sports videos [3].

In Table 4, we provide a comparison of the performance of the techniques *S-MPEG2* and *S-MPEG4-SP*. Results have been obtained under the hypothesis of ideal (error free) annotation engine (events and objects are detected manually), from DV source videos. According with the weights assigned to user, we select different compression factors for objects and events of interests w.r.t. to non interesting elements.

The average PSNR is calculated at fixed bandwidth. In order to maintain the frame rate of 10 fps, and comparable viewing quality, compressed outputs have been obtained at the average bandwidth of 530 kbps for MPEG-2 based solutions, and 180 kbps for MPEG-4-based solutions (please note that MPEG-4 based solutions achieve similar viewing quality with less bandwidth). The average PSNR improvement with semantic adaptation is about 8.5%. Results have been obtained with a user profile of reference (see Table 5). Videos included in the test set take into consideration different sources from different broadcasters and different conditions, and they are selected considering the typical average percentage of highlights in a soccer match, as provided by UEFA organization.

4. PERFORMANCE MEASURE FOR ANNOTATION AND ADAPTATION

Let us consider the case of the access to a whole soccer game (90 minutes) from a mobile device connected with GPRS, whose real average bandwidth can be considered 35 kbps. The use of semantic adaptation enables to achieve acceptable quality for significant entities also with this very strict limit. The 90 minutes of videos are downscaled from the PAL format to a 220x176 frame size, with reference to an off-the-shelf latest generation cellular phone (Motorola V525). An example of the quality of a relevant frame is reported in Fig. 1(b) (and zoomed on the player in Fig. 1(d)) that corresponds to about 32.7 dB. If a standard approach is employed, without exploiting the semantics, results are much poorer, as demonstrated in Figs. 1(a) and 1(c), corresponding to about 30.4 dB.

Nevertheless, standard metrics, such as PSNR and BR for the adaptation module and DR and FAR for the annotation engine, present two main drawbacks for our purposes:

- these metrics evaluate the performance of the single module (annotation or adaptation), but not of the integrated system; in particular, our proposal aims at evaluating how much the annotation errors affects the overall performance of the system;
- these metrics do not take the user's preferences into account; for instance, degrading the quality of different parts of the video can have different impacts on the user, depending on the relevance that those parts have for her/him; standard PSNR does not consider this;

The errors of automatic annotation can affect the user's satisfaction. Since objects and events are divided in classes of relevance by the users, errors can cause under- or over-estimations of objects or events. In particular, *under-estimation* and *miss* conditions have a negative impact on user's satisfaction under the viewpoint of *viewing quality loss*. In fact, in this case, events and/or objects are compressed more than necessary. Instead, costs paid by the user are lowered since under-estimated objects and events are more compressed. On the other hand, *over-estimation* and *false detection* conditions affect negatively user's satisfaction with respect to the *cost* paid by the user (for transmission, downloading, and storage). These two effects could compensate each other: two videos differently annotated could be compressed with the same PSNR and the same BR, but with a large negative impact in user's satisfaction: user can lose details of interests and waste bits for useless parts.

Starting from the usual figures of PSNR at the pixel level and Bit Rate, we can derive new indexes of performance that do not take into account the parts correctly annotated and adapted but only the errors: *i) Viewing Quality Loss (VQL)*: resulting from over-compression due to under-estimation and miss conditions occurred in the annotation; *ii) Bitrate Cost Increase (BCI)*: resulting from higher bitrate due to over-estimations and false detections. Let us call Err^t_Q the set of points of frame t that have been under-estimated, i.e. all the points that are supposed to belong to a class C_i and are, instead, detected as belonging to a class C_j, with $j < i$. Correspondingly, let us call Err^t_C the set of points of frames that have been over-estimated.

The VQL is evaluated on the pixels that result under-estimated for each frame I^t. Using the standard PSNR definition on this set Err^t_Q, a comparison between ideal (error-free) and actual annotation is provided. The PSNR of under-estimated pixels in the case

Profile	$C2$	$C1$	$C0$	**Weights** $(w_{C_2}, w_{C_1}, w_{C_0})$
Profile Ref.	$<$ {SG,FL}, * $>$	$<$ {PK,AA}, players $>$	residuals	(1.0, 0.3, 0.1)
Profile A	$<$ SG, * $>$	$<$ {FL, PK, AA}, players $>$	residuals	(1.0, 0.3, 0.1)
Profile B	$<$ {SG,FL}, * $>$	$<$ {PK, AA}, players $>$	residuals	(1.0, 0.6, 0.5)
Profile C	$<$ {SG, FL}, players $>$	$<$ {PK, AA}, players $>$	residulas	(1.0, 0.3, 0.1)
Profile D	$<$ {SG, FL}, {playfield, players} $>$	$<$ {PK, AA}, players $>$	residuals	(1.0, 0.3, 0.1)

Table 5: User profiles used to evaluate average VQL and BCI

of actual annotation is denoted by $PSNR_{Err^t_Q}$, and defined as:

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

where 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 of the frame (limited to Err^t_Q), defined as follows:

$$MSE_{Err^t_Q} = \frac{\sum\limits_{p\in Err^t_Q} d^2(p)}{|Err^t_Q|} \tag{4}$$

with $d(p)$ a properly defined distance to measure the error between original and distorted images. As distance, we used the *Euclidean distance* in the RGB color space.

The same measure of Eq. 3 can be carried out for ideal (groundtruthed) annotation: the computed $PSNR^{ID}_{Err^t_Q}$ is also computed on the set Err^t_Q and it is is affected by a non null $MSE^{ID}_{Err^t_Q}$, only due to the selected compression standard and the quantization scale. The viewing quality loss of the frame is, thus, defined as:

$$VQL^t = 1 - \frac{PSNR_{Err^t_Q}}{PSNR^{ID}_{Err^t_Q}} \tag{5}$$

Since $PSNR_{Err^t_Q}$ is computed only on under-estimated pixels of the frame, its value is lower or equal to that in the case of error-free annotation ($PSNR^{ID}_{Err^t_Q}$). Consequently, the ratio of Eq. 5 is between 1 (ideal annotation) and 0 (maximum distortion due to annotation and adaptation processes).

Similarly, the *bitrate cost increase*, for objects and events, is defined for a frame I^t as the ratio between the bandwidth request in the ideal and actual case computed on the set of over-estimated pixels Err^t_C:

$$BCI^t = 1 - \frac{BR^{ID}_{Err^t_C}}{BR_{Err^t_C}} \tag{6}$$

Viewing quality loss (VQL) and *bitrate cost increase* (BCI) at the video level are directly obtained by averaging the VQL^t and the BCI^t:

$$VQL = \frac{\sum\limits_{t=0}^{N} VQL^t}{N} \quad ; \quad BCI = \frac{\sum\limits_{t=0}^{N} BCI^t}{N} \tag{7}$$

where N is equal to the number of frames of true highlight plus the number of frames falsely detected.

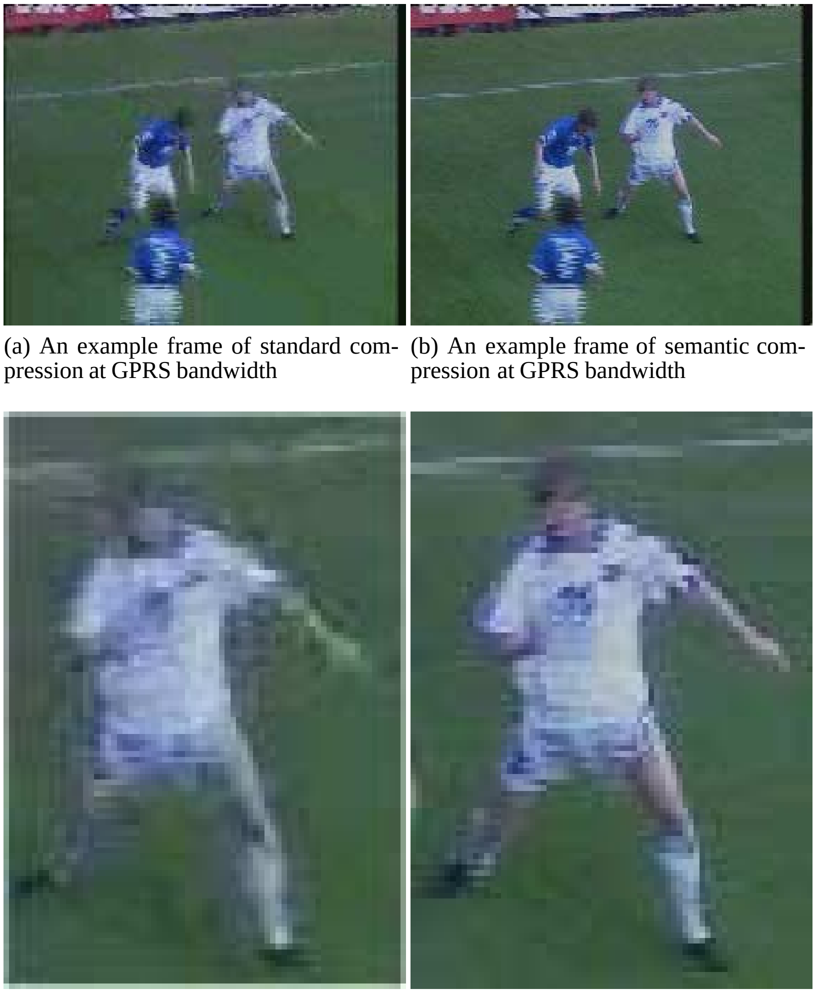
The graphs reported in Fig. 2 compare the performance analysis achievable with classical metrics (such as PSNR and Bit Rate) and with the new metrics, VQL and BCI, for a sample case. The reported example presents a set of annotation errors. The ideal event annotation should detect a "forward launch" (FL) event (associated to class of relevance C_1) between frames 0 and 42, and a "shot on goal"(SG) event (of class C_2) between frames 282 and 375. Annotation with the actual system results in a FL detected between frames 12 and 26, leading to two partial misses (represented by frame intervals 1 and 3 in Fig. 2), and in a SG between frames 251 and 322, resulting in a partial false detection (number 4) and a partial missed detection (number 6). In addition to the errors in event detection, the actual system makes some errors in the segmentation of the objects (the playfield in this case) that result in the small cost increase in intervals 2 and 5, and in more relevant (especially in interval 5) loss of viewing quality. Please note that the descending PSNR in intervals 4 and 5 is due to the decreasing area of the playfield, since the amount of playfield in the image decreases while approaching to the goal. It is also worth noting that the effect of missed event in the case of FL (intervals 1 and 3) and in that of SG (interval 6) is different, being different the relevance of the missed event. The false event in interval 4 results in a BCI of about 80%. In fact, in this interval, the average occupation of a frame in the case of correct annotation is about 100 Kbits, that grows to 650 Kbits since the actual system misclassifies the frame.

From the graphs of Fig. 2 it is evident that the use of classic metrics is not sufficient. The PSNR reported in the upper graph is computed to the whole frame I^t and can mix both quality and cost effects of incorrect annotation. As a limit case, these two effects can neutralize each other, resulting in two videos with the same average PSNR, but very different user satisfaction levels. From PSNR and BR only it is not possible, for instance, to understand how much of the PSNR's decrease of interval 5 is due to annotation errors and how much is due to reduced playfield size.

According to the definitions above, both viewing quality loss and bitrate cost increase depend on the statistics of the objects and events present in the video, the performance of the annotation (measured in terms of misses, misclassifications and losses), the performance of the adaptation, and, ultimately, the way in which objects and events are clustered into each class of relevance and the relative importance weight. Thus it appears that the most important conditions potentially influencing user satisfaction are those related to events. However players objects are also important for their impact on viewing quality, especially in the presence of meaningful actions. Among event conditions the most critical one for performance is event miss. In fact, in this case all the frames during the whole duration of the event are compressed at a lower rate, that is proportional to the relevance weight of the *residual* C_0 class that comprises the events that are less interesting for the user.

5. PERFORMANCE EVALUATION

To evaluate how VQL and BCI change according to different sets of user preferences (CoR and weights) and to the performance of the automatic annotation system (events and objects misses, and wrongly recognized highlights) selected soccer videos taken from the video test set have been manually annotated, performing objects and events segmentation, to precisely estimate under- and over-estimation of objects and events. A reference user profile has been compared versus 4 other possible user profiles, defined as reported in Table 5.



(a) An example frame of standard compression at GPRS bandwidth

(b) An example frame of semantic compression at GPRS bandwidth

(c) Zoomed portion of (a)

(d) Zoomed portion of (b)

Figure 1: Examples of standard compression compared with the semantic approach.

	Ref.	Profile A	Profile B	Profile C	Profile D
FL	4974.95	711.21	4994.88	1222.39	3904.62
SG	5497.50	5497.50	5502.80	1476.17	3683.73
AA	598.89	598.89	797.74	598.89	598.89
PK	485.72	485.72	713.20	485.72	485.72

Table 6: Bitrate of the video obtained with the actual annotation (figures are in kbps).

	Ref.	Profile A	Profile B	Profile C	Profile D
FL	35.47	31.92	35.47	31.85	32.55
SG	33.64	33.64	33.64	31.35	32.39
AA	32.29	32.29	34.01	32.29	32.29
PK	30.88	30.88	32.57	30.88	30.88

Table 7: Peak Signal-to-Noise Ratio (PSNR) of the video obtained with the actual annotation (figures are in dB).

The test set consists in a set of selected clips from different soccer videos: in total about 6000 frames have been annotated with 2361 highlights frames (624 frames of forward launches, 760 of shot on goal, 320 of attack actions, 650 of placed kicks). Also the annotation at object level has been provided with players and playfield. Both manual and annotated versions have been adapted with $S - MPEG2$, according with the classes of relevances of the 5 users. $S - MPEG2$ has been preferred to $S - MPEG4 - SP$ for two main reasons: first, because MPEG-2 is less computational intensive, and, thus, more suitable for low-power devices; second, as stated above, MPEG-2 enables selective compression at both object and event level.

The content-based adaptation system provides a compression at constant quality for the pixels belonging to the highest classes of relevance. For this reason, similarly to the Constant Quality (CQ) method of MPEG, we define this approach as *Constant Best Quality* (CBQ). Tables 6 and 7 reports the average values divided for type of highlights. For instance, User Ref and User A are different only regarding FL. User B has the same PSNR and BR for the FL and SG highlights (both of class C_2), while obtains a higher overall quality (PSNR) than User Ref in AA and PK highlights (both of class C_1). User C and User D have apparently lower PSNR than Ref because it is averaged over all the pixels of the frame. These profiles, indeed, ask for the best quality only for the regions of interest (players and playfield, respectively). The video quality in interesting areas is almost same, but a decrease in the required bandwidth is evident.

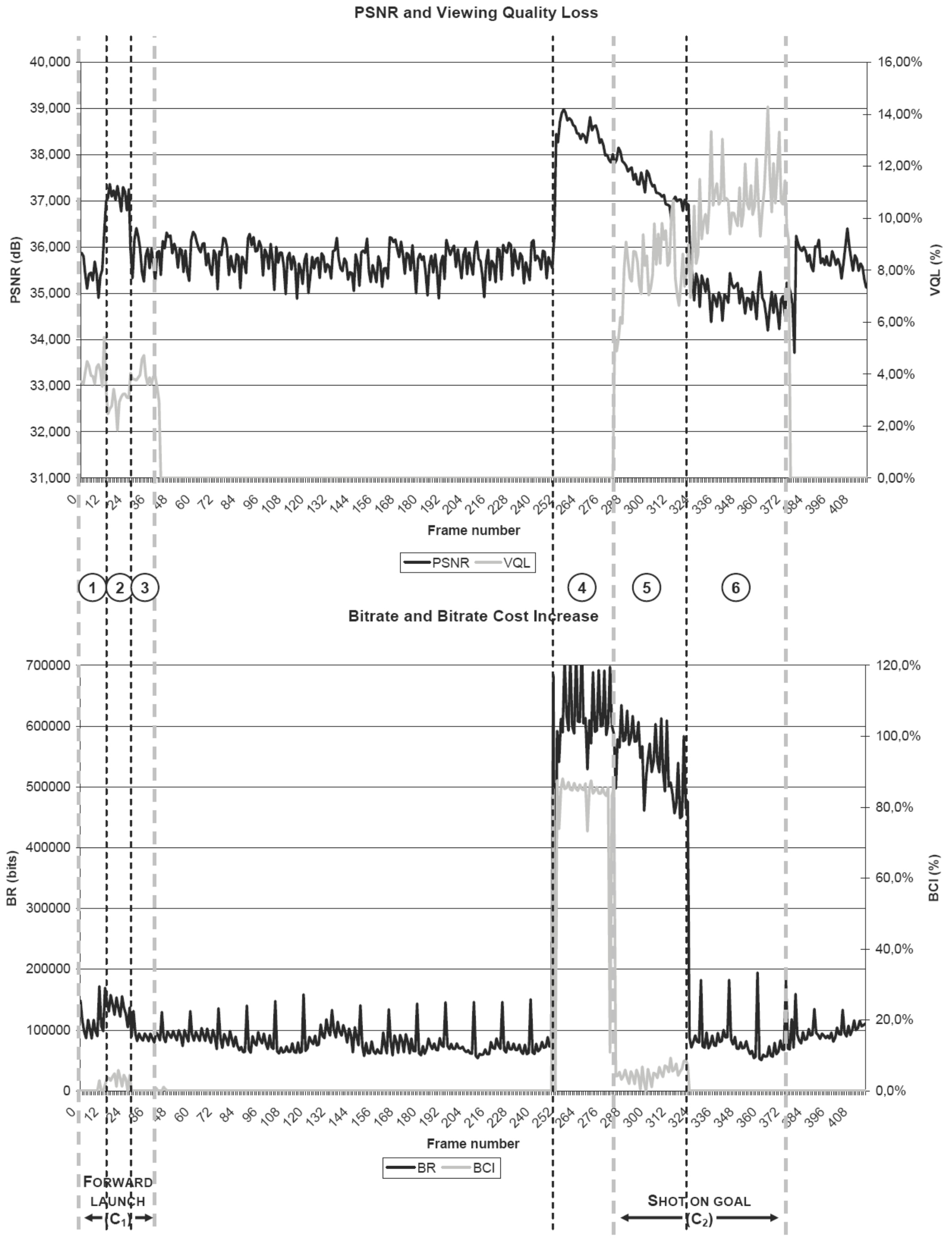


Figure 2: Comparison between PSNR-BR classical metrics and newly defined VQL and BCI.

Finally, Tables 8 and 9 report the results achieved with the new metrics. Here, many considerations about the goodness of the whole annotation and adaptation systems can be inferred. First, there is a high bitrate cost increase due to the errors in shot of goal. This is due to the high number of false positives and the average length of shot on goal, i.e. to the number of frames erroneously classified as SG. It is worth noting that, from Table 2, FL presents also higher false alarm rate at event level than SG, but the BCI is always lower than that of SG. This can be explained because shots on goal are highlight that last more frames, thus the over-estimated frames in the case of missed events are more (the average length for shot on goal is about 140 frames, while in the case of forward launch is 60 frames). Another interesting results provided by the BCI measure is that obtained comparing User Ref and User D. It can be easily noted that, at least for FL and SG, the BCI is higher for User D. The BCI is due to two factors: false/over-estimated events and false/over-estimated objects. In the first case, the BCI is similar to that of User Ref since the playfield is usually very large in SG and

	Ref.	Profile A	Profile B	Profile C	Profile D
FL	9.07%	1.23%	8.14%	5.58%	9.39%
SG	11.78%	11.78%	11.07%	8.10%	13.76%
AA	0.45%	0.45%	0.56%	0.45%	0.45%
PK	0.27%	0.27%	0.30%	0.27%	0.27%

Table 8: Bitrate Cost Increase.

	Ref.	Profile A	Profile B	Profile C	Profile D
FL	2.75%	0.61%	1.47%	2.49%	3.60%
SG	1.31%	1.31%	0.83%	3.04%	2.54%
AA	1.34%	1.34%	1.14%	1.34%	1.34%
PK	0.24%	0.24%	0.24%	0.24%	0.24%

Table 9: Viewing Quality Loss.

FL actions. In addition, in the case of User D, there is the overestimation of the playfield that contributes to increasing the BCI. A similar consideration can be done for User C, but in this case the of missed/over-estimated events the errors is lower since the objects (players) are smaller and the number of over-estimated pixels smaller. Thus, the overall BCI is lower than in the case of User Ref. and User D.

Regarding VQL, the error in quality is limited, especially for User B. In fact, the User B accepts higher costs (due to a higher bitrate) that limit the effects of miss detections. User A that is less interested in FL is not affected by significant errors. AA and PK highlights are always considered of average importance.

Acknowledgments

The work has been carried out in the context of DELOS, the Network of Excellence in digital libraries of European VI Framework programme.

6. REFERENCES

[1] 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.

[2] M. Bertini, R. Cucchiara, A. D. Bimbo, and A. Prati. An integrated framework for semantic annotation and transcoding. *Multimedia tools and applications*, to appear.

[3] M. Bertini, R. Cucchiara, A. Del Bimbo, and A. Prati. Object-based and event-based semantic video adaptation. In *Proceedings of Int'l Conference on Pattern Recognition, to appear*, Aug. 2004.

[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] 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.

[7] M. Kunt. *Object-based Video Coding*, chapter 6.3, pages 585–596. in 'Handbook of Image and Video Processing'. Academic Press, 2000.

[8] R. Mohan, J. Smith, and C. Li. Adapting multimedia internet content for universal access. *IEEE Transactions on Multimedia*, 1(1):104–114, March 1999.

[9] T. Shanableh and M. Ghanbari. Heterogeneous video transcoding to lower spatio-temporal resolution and different encoding formats. *IEEE Transactions on Multimedia*, 2(2):101–110, June 2000.

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

[11] B. L. Tseng, C.-Y. Lin, and J. R. Smith. Using MPEG-7 and MPEG-21 for personalizing video. *IEEE Multimedia*, 11(1):42–52, Jan.-Mar. 2004.

[12] A. Vetro. MPEG-21 digital item adaptation: enabling universal multimedia access. *IEEE Multimedia*, 11(1):84–87, Jan.-Mar. 2004.

[13] A. Vetro, C. Chrisopoulos, and H. Sun. Video transcoding architectures and techniques: An overview. *IEEE Signal Processing Magazine*, 20(2):18–29, Mar. 2003.

[14] 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.

[15] Y. Wang, J.-G. Kim, and S.-F. Chang;. Content-based utility function prediction for real-time MPEG-4 video transcoding. In *Proc. of IEEE Int'l Conference on Image Processing*, volume 1, pages 189–192, 2003.

[16] W. Zhou, A. Vellaikal, , and C. Kuo. Rule-based video classification system for basketball video indexing. In *Proc. ACM Multimedia 2000 workshop*, pages 213–216, 2000.