Content-based Video Adaptation with User's Preferences

M. Bertini†, R. Cucchiara‡, A. Del Bimbo†, A. Prati‡
†*D.S.I. - Universitá di Firenze - Italy* ‡*D.I.I. - Universitá di Modena e Reggio Emilia - Italy*

Abstract

In this paper, we present an integrated system that has been designed to support automatic semantic extraction of highlights in sports video and automatic video adaptation according to user's preferences. To analyze the user's satisfaction, we propose a new performance measure that explicitly takes into account the user's preferences and considers the number and type of errors produced by the annotation engine and the way in which these errors affect the compressed video quality and bandwidth allocation. We provide experimental results with application to soccer and swimming.

1. Introduction

The availability of reliable connections through handheld mobile devices such as PDAs, HCCs or smart phones has introduced the possibility of accessing multimedia information from remote clients at any place. Limitations are mainly concerned with the cost of data transfer that is paid by the user (which is directly connected to the bandwidth available and, hence, to the quality of transmission) and to the limited displaying capability of the device and its portability (that determine that it should be used primarily for short time lags, in a non-continuous way). A typical service is the transmission (possibly with high viewing quality) of meaningful highlights extracted from a continuous video stream, usually at the occurrence of significant events.

In this framework, system services must be personalized to a specific user profile. We expect that the user can elicit relevant entities (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 and the quality of their transmission should be adapted to their relevance. Video adaptation in terms of the relevance of the objects detected in each frame has been addressed by [6] and [2], for video surveillance applications. Chang et al. [3] have filtered live video content according to events and highlights. Bertini et al. [2] have developed a prototype system for annotation and adaptation of soccer sport videos, with adaptation based on objects and highlights. Related works in the literature have commonly evaluated the performance of content-based video adaptation systems in terms of efficacy. The Open Video Project [4] has identified the speed of search and the subjective recall,as the basic parameters for the evaluation of user's satisfaction in video retrieval on demand. The PSNR is used instead as a measure of the viewing quality of compressed video, even if more sophisticated measures accounting for non-linear distortion effects on the human perception system could be used [6, 5]. A weighted PSNR has been defined in [2] to include user's preferences. Chang et al. [3] 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 all these experiences, both the viewing quality and the bit rate are not put in relationship with the user's interest, nor with the capability of the system to detect automatically user-relevant elements.

In this paper we present an integrated system that supports automatic semantic annotation and adaptation of sports video. The adaptation is based on objects and highlights, according to user-defined *classes of relevance*. We also propose a new model to measure the user's satisfaction that combines a quantitative measure of compressed video quality and used bandwidth, and also takes into account the compliance with the user's preferences, considering the number and type of errors produced by the automatic detection subsystem and the way in which these errors affect the compression. We provide experimental evidence and measures of user's satisfaction with application to soccer and swimming.

2. Content-based adaptation with automatic extraction of meaningful entities

The scheme reported in Fig. 1 shows the basic principles of operation of the overall system. The semantic annotation engine [1] extracts from the raw video the meaningful objects (o_i) and events (e_i). The system performs automatic detection of most prominent highlights in different sports. Low-level features are used to detect the playfield zones, the camera motion and players' positions. Objects that are detected automatically are the playfield zones,

the players' blobs and the background. Events are modeled with Finite State Machines, where combinations of feature values determine the transition from one state to the following. Event models are checked against the current observations through a model checking algorithm.

Objects and events are assigned to classes of relevance (C_i). A class of relevance is defined as a set of meaningful elements (objects and events) that the user is interested in and the system is able to manage. In this way, the user can assign a degree of preference to each class, in order to have the best quality/cost trade-off for the most relevant classes at the price of a lower quality for the least relevant ones.

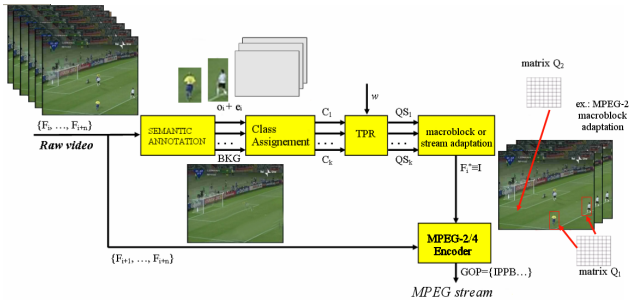


Figure 1. General scheme of the proposed system

The Transcoding Policy Resolver engine (TPR) selectively compresses the objects and events detected by the annotation engine following the specifications indicated in their class of relevance. Two different coding policies have been implemented that perform coding at the semantic level: i) a modified version of MPEG-2 (referred to as *S-MPEG2*) - where compression is applied to sets of pixels in each frame that identify objects - in which the quantization factor QS_i of each macroblock i of each frame is chosen according to the dominant class of relevance that has been assigned to the macroblock, depending on which objects and event are involved; ii) a modified version of MPEG-4, Simple profile (referred to as *S-MPEG4*), - where compression is applied to sequences of frames that identify events - in which the compression factor that is applied to each frame depends on the class of relevance that has been assigned to the clip which the frame belongs to.

S-MPEG2 normally outperforms standard MPEG-2 but has lower performance than standard MPEG-4. However, due to the simplicity of the MPEG-2 decoder it is easier to implement than MPEG-4, with simple PDAs with Windows CE operating system. On the other hand, the standard MPEG-4 introduces a sensible overhead in handling the object's alpha planes and is suitable only in conditions where almost fixed background is present. This motivates the use of S-MPEG4 with semantic encoding only at the event level, since in S-MPEG4 the Simple profile limits the range of variability of the quantization factor of the macroblocks of a single frame [2].

Case study	MPEG-2		MPEG-4	
	Standard	S-MPEG2	Standard	S-MPEG4
Soccer	34,13 dB	35,93 dB	33,38 dB	37,30 dB
Swimming	30,56 dB	32,99 dB	30,70 dB	33,27 dB

Table 1. Average PSNR for ideal annotation.

The two semantic coding systems have been tested over about 1 hour of soccer video and 25 minutes of swimming videos. Table 1 reports the average PSNR for S-MPEG2 and S-MPEG4 w.r.t. MPEG-2 and MPEG-4 standards with no semantic encoding. Results are reported for comparable bandwidths within the same compression standard. The same bit rate of a typical GPRS bandwidth is used, for both semantic and non semantic coding; however, typically, MPEG-4 bandwidth is lower (one quarter) than that of the MPEG-2, although PSNR is similar or even higher, in most cases. Results of S-MPEG2 and S-MPEG4 have been obtained under the hypothesis of ideal (error free) annotation subsystem (events and objects detected manually).

Actually, the semantic annotation engine introduces errors that are typically measured in terms of "detection rate" (DR) or "recall", and "false alarm rate" (FAR) or "precision" ($1 - FAR$). Table 2 reports these measures for the detection of meaningful highlights in soccer and swimming for the same test set. Measures are provided at the event-level (the capability to detect correctly the occurrence of a certain event in the video stream), at frame-level (the capability to assign to each individual frame the correct event to which it belongs to), and at object-level (the capability to detect correctly the occurrence of a certain object within individual frames). Typically, measures at the frame-level have a lower DR and higher FAR than those at event-level. Measures at the frame and object level are those important for content-based video adaptation, as they directly affect the user's satisfaction.

3. Performance measures accounting for user's preferences and satisfaction

Measures in Table 1 and 2 provide absolute performance figures. Since they are derived independently from the user's preferences, they are not useful to understand the degree of satisfaction of the user in practical applications. Instead *from the user point of view* errors that occur in the annotation subsystem result in an under- or over-estimation of objects or events. Considering events and objects, we can distinguish the following cases:

$< E_c, E_u, E_m, E_o, E_f >, < O_c, O_u, O_m, O_o, O_f >$,

where E_c, is for a correctly classified event; E_u, is for an under-estimated event (event assigned to a less impor-

Sports video	Highlight	Event-level		Frame-level		Object-level	
		DR	FAR	DR	FAR	DR	FAR
Soccer videos	SG	77,78 %	6,67%	67,23%	34,62%	99,94%	0,16%
	PK	100,00%	13,33 %	100,00%	14,21%		
	FL	88,89 %	3,03 %	63,33%	7,74%		
Swimming videos	ST	71,43 %	16,67 %	70,91%	0,00 %	99,85%	1,01%
	TU	100,00 %	20,00 %	86,23 %	6,30 %		
	AR	85,71 %	14,29 %	77,50 %	39,81 %		

Table 2. Performance figures of the semantic annotation engine. Highlights detected are Shot on goal (SG), Placed Kick (PK), and Forward Launch (FL) - for soccer - and Start (ST) Turning (TU) and Arrival (AR) - for swimming. The Detected object is the playfield in both sports.

tant class of relevance than expected); E_m is for a missed event (event not recognized as such); E_o, is for an over-estimated event (event assigned to a more important class of relevance than expected); E_f, is for a falsely detected event (something not relevant that is considered as relevant). The cases for O_c, O_u, O_m, O_o, and O_f referred to objects, are defined similarly.

Different types of errors affect user's satisfaction differently. In particular *under-estimation* and *miss* conditions (E_u, E_m, O_u, O_m) have a negative impact on user's satisfaction under the viewpoint of *viewing quality loss*. Costs paid by the user are instead lowered since under-estimated objects and events are more compressed. On the other hand, *over-estimation* and *false detection* conditions (E_o, E_f, O_o, O_f) affect negatively user's satisfaction with respect to the *cost* paid by the user (for transmission, downloading and storage). In fact, in these cases, non interesting parts are classified as relevant, and are produced at a higher viewing quality, thus having a cost higher than expected.

Considering these two effects, the error cases can be combined to form the following categories of user's unsatisfaction sources (in order of descending user's satisfaction provoked):

$$NoErr = E_c \wedge O_c$$
$$Err_{Q_o} = E_c \wedge (O_u \vee O_m) \quad Err_{C_o} = E_c \wedge (O_o \vee O_f)$$
$$Err_{Q_e} = E_u \vee E_m \quad Err_{C_e} = E_o \vee E_f$$

The $NoErr$ case reflects the ideal case (no errors) for the annotation subsystem; user's unsatisfaction is eventually related to the unavoidable compression of the adaptation module (see Table 1). Given a pixel p, we define $MSE_{NoErr}(p)$ the measure of the distortion introduced by the content adaptation subsystem and we can normalize the other measures with respect to it. Then, the *quality error rate* (viewing quality loss) for objects and events can be defined as:

$$\epsilon_{Q_o} = 1 - \frac{MSE_{NoErr}}{MSE_{Err_{Q_o}}} \quad ; \quad \epsilon_{Q_e} = 1 - \frac{MSE_{NoErr}}{MSE_{Err_{Q_e}}} \quad (1)$$

The denominator values are the measured MSE in the case of Err_{Q_o} and Err_{Q_e}. The ratio will range between 1 (ideal annotation) and 0 (maximum distortion due to annotation and adaptation processes). The quality error rate at the pixel level can be integrated to consider all pixels p of the frame that are assigned to a certain class of relevance C_i:

$$QErr^{frame}_{C_i} = \frac{\sum_{p \in C_i} (\epsilon_{Q_o}(p) + \epsilon_{Q_e}(p))}{|C_i|} \quad (2)$$

At the frame level, $QErr^{frame}$ is obtained summing up all the $QErr^{frame}_{C_i}$ for all the classes of relevance, each weighted according to the weight w_i that has been assigned to it:

$$QErr^{frame} = \sum_{i=0}^{N_{CL}} w_i QErr^{frame}_{C_i} \quad (3)$$

Similarly, the *bandwidth waste* is computed, directly at frame level, as the ratio between the bandwidth in the case $NoErr$ and that in the cases Err_{C_o} and Err_{C_e}:

$$\epsilon_{C_o} = 1 - \frac{BR_{NoErr}}{BR_{Err_{C_o}}} \quad ; \quad \epsilon_{C_e} = 1 - \frac{BR_{NoErr}}{BR_{Err_{C_e}}} \quad (4)$$

A global measure is obtained by summing up these two factors.

$$CErr^{frame} = \epsilon_{C_o} + \epsilon_{C_e} \quad (5)$$

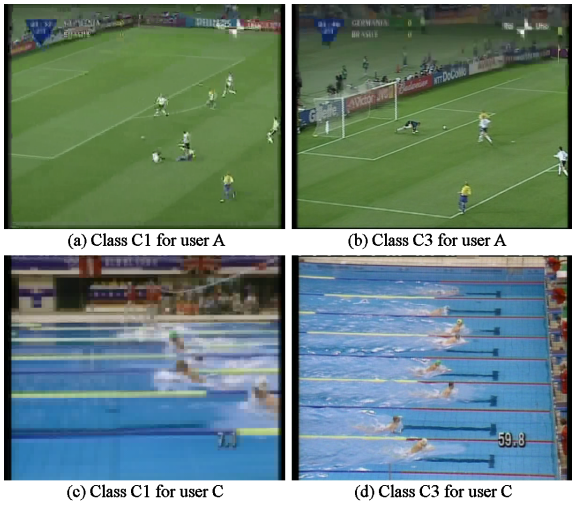
Finally , measures of viewing quality loss and bandwidth waste, $QErr$ and $CErr$, at the video level are obtained by averaging the $QErr^{frame}$ and the $CErr^{frame}$ over all the frames of the video.

4. System Performance Analysis

To evaluate the performance of our system for automatic sports video annotation and adaptation, we have created a ground truth by segmenting manually events (highlights) and objects (playground, crowd, players) in sports videos of different sports. The system has then been tested under different profiles of user's satisfaction, defined through appropriate classes of relevance of events and objects. The

Soccer videos						Swimming videos		
User profile A			User profile B			User profile C		
	$QErr$	$CErr$		$QErr$	$CErr$		$QErr$	$CErr$
C_1	n/a	3,09%	C_1	n/a	3,09%	C_1	n/a	2,07%
C_2	1,80%	0,03%	C_2	3,24%	n/a	C_2	1,34%	0,06%
C_3	2,62%	n/a				C_3	1,90%	n/a

Table 3. Performance measures affecting user's satisfaction for the different classes of relevance and with the user's profiles A, B, and C.



(a) Class C1 for user A (b) Class C3 for user A

(c) Class C1 for user C (d) Class C3 for user C

Figure 2. Examples of viewing quality loss (user profile A and C).

S-MPEG2 has been adopted as transcoding policy for the tests. Using the performance indexes for viewing quality and cost increase defined in the previous section, we can provide a measure of the system performance that indicates how the system operates with respect to user's preferences and can be used to guarantee a degree of satisfaction to the final user for viewing quality and cost increase. Average measures of the system performance with these indexes have been reported in Table 3 with reference to three different user profiles. Values have been obtained from the same test set of about 1 hour of soccer video and 25 minutes of swimming video.

User profile A is interested in soccer videos, and has defined the following classes of relevance: C_1 = any other object/event not included in C_2 and C_3; C_2 = $playfield$ object and $forward\ launch$ event; C_3 = $playfield$ object and $shot on goal$ or $placed kick$ event. User profile B is also interested in soccer, but with different preferences. His/her classes of relevance are: C_1 = any other object/event not included in C_2; C_2 = $playfield$ object and $forward\ launch$ or $shot\ on\ goal$ or $placed\ kick$ event. User profile C is interested in swimming, and has defined three classes of relevance: C_1 = any other object/event not included in C_2 and C_3; C_2 = $pool$ object and $turning$ event; C_3 = $pool$ object and $start$ or $arrival$ event. Fig. 2 shows examples of the viewing quality associated with events in class C_1 and C_3 for two of the above defined user profiles (A and C).

In soccer videos, we can see that user profile B, who is interested in viewing at high quality any highlight, has actually 3.24% of viewing quality loss. In addition, a bandwidth waste of 3.09% is observed. We can also observe that user profile A, who is interested in viewing at high quality only a subset of the highlights requested by profile B has the same costs of profile B, in that false detections of the *forward launch* highlight (that is assigned to a less relevant class of relevance) are almost ininfluent, as from Table 2.

References

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

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

[3] S. Chang, D. Anastassiou, A. Eleftheriadis, J. Meng, S. Paek, S. Pajhan, and J.R. Smith. Content-based video summarization and adaptation for ubiquitous media access. In *Proc. of Int'l Conference on Image Analysis and Processing*, pages 494–496, September 2003.

[4] Gary Geisler and Gary Marchionini. The open video project: research-oriented digital video repository. In *ACM DL*, pages 258–259, 2000.

[5] Wilson S. Geisler Brian L. Evans Niranjan Damera-Venkata, Thomas D. Kite and Alan C. Bovik. Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9(4):636–650, April 2000.

[6] A. Vetro, T. Haga, K. Sumi, and Huifang Sun;. Object-based coding for long-term archive of surveillance video. In *Proceedings of International Conference on Multimedia & Expo (ICME)*, volume 2, pages 417–420, 2003.