Object-based and Event-based Semantic Video Adaptation

M. Bertini, A. Del Bimbo
Dip. di Sistemi e Informatica
University of Florence

R. Cucchiara, A. Prati
Dip. di Ingegneria dell'Informazione
University of Modena and Reggio Emilia

Abstract

Semantic video adaptation allows to transmit video content with different viewing quality, depending on the relevance of the content from the user's viewpoint. To this end, an automatic annotation subsystem must be employed that automatically detect relevant objects and events in the video stream. In this paper we present a composite framework that is made of an automatic annotation engine and a semantics-based adaptation module. Three new different compression solutions are proposed that work at the object or event level. Their performance is compared according to a new measure that takes into account the user's satisfaction and the effects on it of the errors in the annotation module.

1. Introduction and Related Work

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.

Video adaptation techniques have been widely studied in the last years [7, 5] in order to meet the constraints of the limited resources of the devices, to satisfy the user's requirements, and, at the same time, to keep low the costs of the transmission in terms of data transferred and time required. Most of the video adaptation techniques provide syntactic video adaptation performing scaling, color subsampling, temporal downscaling or changing the compression's factor [6]. 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 [8] and [2] for video surveillance applications. In [8], Vetro et al. presented an object-based transcoding framework that uses dynamic programming or meta-data, for the allocation of bits among the multiple objects in the scene. Chang et al. [3] have filtered live video content according to highlights. In [2] authors have developed a prototype 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.

Most of measures for performance evaluation of video annotation systems are, however, still based on the PSNR (Peak Signal-to-Noise Ratio) [3, 2] with some noticeable exceptions that take into account non-linear distortion effects on the human perception system [8, 4]. 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 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. [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).

The contribution of this paper is twofold. First, we propose content-based video adaptation techniques at object- and event-level based on suitable modifications of MPEG-2 and MPEG-4 standards. Then, a novel 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 is defined. Effects of object- and event-level adaptation techniques are compared according to this new metric with reference to sports (soccer and swimming) video.

2. Object-based versus Event-based

The content-based video adaptation system results from the integration of an automatic annotation engine and a semantic adaptation module. The annotation engine [1] extracts from the raw video the meaningful objects (o_i) and events (e_i). The annotation engine has been targeted to perform automatic detection of prominent highlights in different sports. Low-level features (such as motion fields or lines) are combined to detect the playfield zones, the camera motion and the players' positions. Objects that are detected automatically are the playfield zones, the players' blobs and the background. Events are modelled 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. The system is able to extract highlight, such as shots at goal, free kicks, and forward launches for the soccer, and start, arrival, and turning for the swimming.

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. 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 $QS_i \in [0, 31]$ 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.

The other two coding policies implemented are 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 *S-MPEG4-SP* is a modified version of the MPEG-4 Simple Profile: 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.

The *S-MPEG4-CP* was instead at object-level, exploiting the idea defined in the Core Profile of MPEG-4 and creating a different VOP for each object extracted by the annotation system. In this way, we can assign different quantization scales to each object in dependence to its relevance for the user. The Core Profile is very used for object animation and graphical manipulation, but for video adaptation it exhibits a main drawback: object-based coding is not suited for video in which camera is moving and objects change often their shape and/or position, since temporal prediction can not be deeply exploited (as for the case of sport videos).

Although performance of MPEG-4 techniques are known to be one order of magnitude better than that of MPEG-2 techniques, we reported also these latter approaches since they are computationally less expensive and, thus, more suitable for the porting on embedded operating system for mobile devices.

3. Performance Evaluation with Ideal Annotation

In this section, we will provide a comparison of the performance of the techniques presented in Section 2. Results have been obtained under the hypothesis of ideal (error free) annotation engine (events and objects are detected manually). 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.

Case study	**MPEG-2**		**MPEG-4**	
	Standard	**S-MPEG2**	**Standard**	**S-MPEG4-SP**
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 system has been tested over about 1 hour of soccer videos and 25 minutes of swimming videos. Table 1 reports the average PSNR for *S-MPEG2* and *S-MPEG4-SP* with respect to MPEG-2 and MPEG-4 standards with no semantic encoding. Results are reported for comparable bandwidths within the same compression standard. However, typically, MPEG-4 bandwidth is much lower than that of the MPEG-2, although PSNR is similar or even higher, in most cases.

Fig. 1(a) reports the comparison between *S-MPEG4-SP* and *S-MPEG4-CP* techniques on a sample soccer video, where compression is applied to the playfield object and

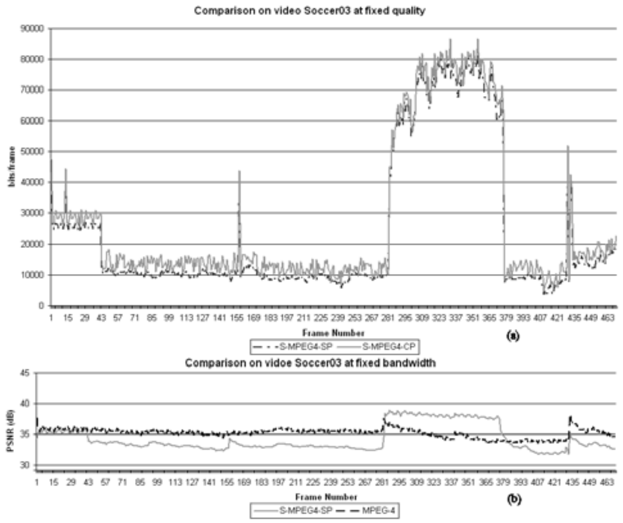

Figure 1. Comparison for a soccer video: (a) between *S-MPEG4-SP* and *S-MPEG4-CP* with fixed quality, and (b) between MPEG-4 and *S-MPEG4-SP* with fixed bandwidth.

to the shot on goal and forward launch events. The comparison is performed by keeping almost the same average PSNR for the two approaches and evaluating the bandwidth frame-by-frame. Between frames 283 and 377 an interesting event occurs and this is transmitted at a higher quality, thus with a larger amount of bits. However, it is possible to note that *S-MPEG4-SP* slightly outperforms *S-MPEG4-CP*. In fact, the average bitrates are 517.18 and 591.30 kbps, respectively. It must be noticed that these results directly depend on the video content. In particular, they can be generalized to the case of moving camera and large objects (like the playfield in sport videos). The reverse can be observed in contexts with fixed camera and small objects with shape changes (like persons in surveillance videos).

The benefits of *S-MPEG4-CP* event-based coding in the case of soccer video are evident in Fig. 1(b) in which *S-MPEG4-SP* technique is compared in terms of PSNR with a MPEG-4 of the same bandwidth. Given that the interesting highlight is between frames 283 and 377, the *S-MPEG4-SP* is able to exploit the information provided by the annotation to increase the quality of that part. As a consequence, in the case of limited availability of bandwidth, instead of compressing equally all the frames of the video (wasting bandwidth for useless parts and reducing quality of interesting ones), we can keep accessible the video and, at the same time, "see better what we want to see".

4. Performance Evaluation with Automatic Annotation

In automatic annotation systems, many errors occur both at object- and at event-level. These errors affect the user's satisfaction due to 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). 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 category of $QErr$ and $CErr$, considering errors affecting quality or cost, respectively. Given a pixel p, we define $MSE_{NoErr}(p)$ the measure of the distortion (in terms of Mean Square Error) introduced by the content adaptation only (as in previous Section), 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:

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

The ratio ranges between 1 (ideal annotation) and 0 (max 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, weighting the different classes of relevance according to the user-defined relevance:

$$QErr^{frame} = \sum_{i=0}^{N_{CL}} w_i \left(\frac{\sum_{p \in C_i} (\epsilon_{Q_o}(p) + \epsilon_{Q_e}(p))}{|C_i|} \right) \tag{2}$$

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}:

$$CErr^{frame} = \left(1 - \frac{BR_{NoErr}}{BR_{Err_{C_o}}}\right) + \left(1 - \frac{BR_{NoErr}}{BR_{Err_{C_e}}}\right) \tag{3}$$

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.

The performance of the *S-MPEG4-SP* techniques for video adaptation used in conjunction with the automatic video annotation system has been evaluated in terms

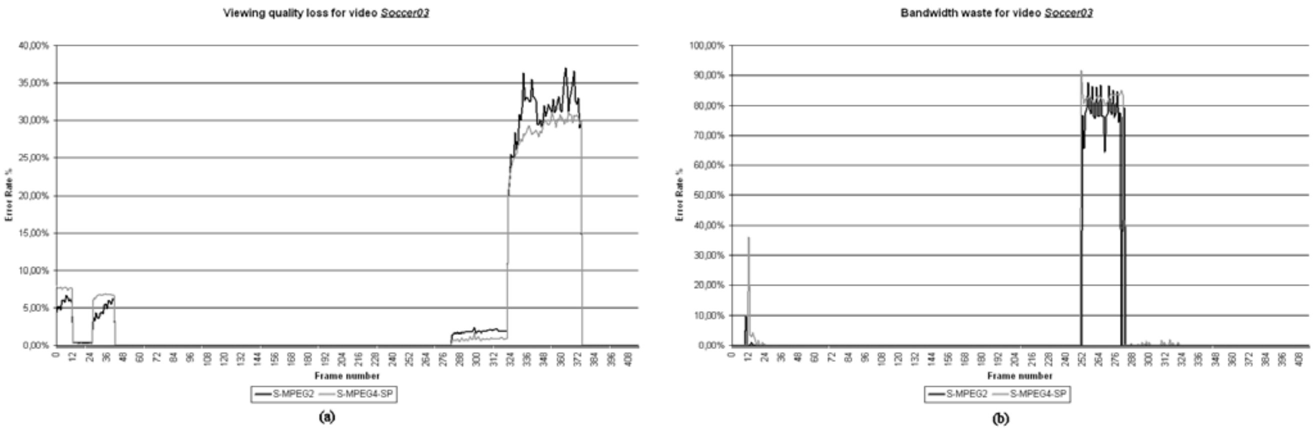

Figure 2. Comparison between *S-MPEG2* and *S-MPEG4-SP* with the new metric.

of user's satisfaction. Fig. 2 compares *S-MPEG2* and *S-MPEG4-SP* for the same sample video of Fig. 1.

The original video presents two highlights: a minor event between frames 0 and 43, and a more important highlight between frames 283 and 377. The annotation engine detects the events between frames 11-25 and 252-323, respectively. This results in under- and over-estimation errors. For example, referring to the second event in Fig. 2(b), the high peak in frames 252-282 is due to the over-estimation of the event and results into a bandwidth waste, whereas the right peak (frames 324-377 in Fig. 2(a)) is due to a missed event and results into viewing quality loss. Between frames 283 and 323 the event is correctly detected, but some errors at object-level are present (small errors in the range of 2-3%).

Nevertheless, we can notice that annotation errors result into comparable viewing quality loss and bandwidth waste in both *S-MPEG2* and *S-MPEG4-SP*. In particular, average viewing quality loss is 4.51% and 4.21% and average bandwidth waste is 5.67% and 6.34% for *S-MPEG2* and *S-MPEG4-SP*, respectively. This demonstrates the robustness of both approaches to annotation errors.

5. Conclusions

In this paper we have presented a composite framework for content-based video adaptation that is made of an automatic annotation engine and a semantic adaptation module, and a novel performance measure that takes into account user's satisfaction and the effects on it of annotation errors. Three new different semantic adaptation policies are presented that perform compression at object- or event-level. Performance comparisons have been reported on sample sport videos that demonstrate that *S-MPEG4-SP* and *S-MPEG4-CP*, respectively working at event- and object-level, both outperform *S-MPEG2*. Moreover, *S-MPEG4-SP* shows better performance than *S-MPEG4-CP* in videos with moving camera and large still objects.

Authors would like to thank Nicola Bicocchi for his help and the European NoE DELOS (contract no. G038-507618) for the financial support.

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] S. Chang, D. Anastassiou, A. Eleftheriadis, J. Meng, S. Paek, S. Pajhan, and J. 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, Sept. 2003.

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

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

[6] 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.

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

[8] A. Vetro, T. Haga, K. Sumi, and H. 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.