Proceedings of IEEE International Conference on Multimedia and Expo, ICME'03, vol. 3, pp. 597-600, Baltimore, USA, July 6-9, 2003.

SEMANTIC SEGMENTATION AND DESCRIPTION FOR VIDEO TRANSCODING

Andrea Cavallaro, Olivier Steiger, Touradj Ebrahimi

Signal Processing Institute
Swiss Federal Institute of Technology (EPFL)
CH-1015 Lausanne, Switzerland

ABSTRACT

We present an automatic content-based video transcoding algorithm which is based on how humans perceive visual information. The transcoder support multiple video objects and their description. First the video is decomposed into meaningful objects through semantic segmentation. Then the transcoder adapts its behaviour to code relevant (foreground) and non relevant objects differently. Both object-based and frame-based encoders are combined with semantic segmentation. Experimental results show that the use of semantics and description prior to transcoding reduces the bandwidth requirements and makes it possible to adapt the video representation to limited network and terminal device capabilities still retaining the essential information.

1. INTRODUCTION

The diffusion of network appliances such as cellular phones, personal digital assistants, and hand-held computers creates a new challenge for content delivery: how to adapt the media transmission to the various device capabilities [1, 2, 3]. Each device is characterized by a certain screen size, color depth, and processing power. Furthermore, such appliances are connected through different kinds of connections with diverse bandwidths. Finally, different users might want to access the same multimedia content. There is therefore the need to adapt the media transmission to network characteristics, terminal capabilities, and user's preferences. The proposed work extends the approach proposed in [5] by considering multiple simultaneous objects and their description (Figure 1). Multiple simultaneous objects are computed through semantic segmentation which identifies portions of the video that are of interest to the user. The semantics is dependent on the application. In our case, we are interested in moving objects and therefore the semantics is defined by motion. Segmentation separates moving objects from each other and from the background. In the specific implementation of this paper, we use the semantic algorithm presented in [4] where video objects are automatically segmented and tracked over time. In addition to this, descriptors are extracted from the video objects and used for transcoding. The details of the proposed method and a comparison with other modalities are presented in Sec. 2 and Sec. 3, respectively. Finally, in Sec. 4 we draw the conclusions.

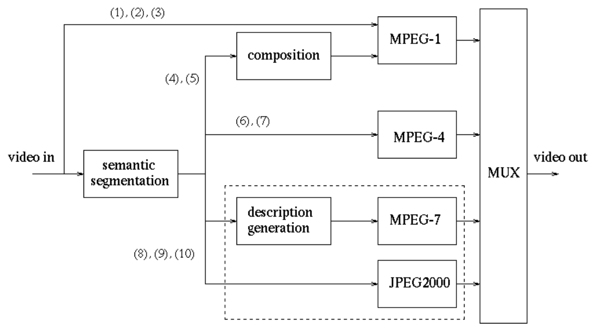

Fig. 1. Flow diagram of the proposed video transcoder based on semantic segmentation and description

2. VIDEO TRANSCODING

Video transcoding is the process of converting video data into another more desirable format. We can identify three main approaches to video transcoding: content-blind transcoding, semantic transcoding, and description-based transcoding. These three modalities are described in the following sections.

2.1. Content-blind transcoding

Traditional transcoding techniques do not perform any semantic analysis of the content prior to conversion. The choice of the output format is determined by network and appliance constraints, independent of the video content (i.e. independent of the way humans perceive visual information).

We can identify three main traditional or content-blind transcoding categories, namely spatial conversion, temporal conversion, and color-depth reduction. *Spatial conversion* affects the height and the width of each frame. A reduction of the frame size reduces the bandwidth requirements

(Fig. 2(c)). *Temporal conversion* modifies the frame rate. Frame rate reduction is acceptable in case the motion activity in the video is limited. When the motion activity is high, temporal conversion may signify the loss of the impression of motion continuity for the user, thus sensibly reducing the perceived quality. A *reduced color-depth* can be obtained when the color quantization step is enlarged so as to reduce the amount of colors to code. This solution is used to take advantage of the known limitations of display capabilities of the user's terminal to reduce the bandwidth. A further simplification introduced by the above mentioned process is the use of grayscale video. Each pixel is described with a gray level only instead of with a three dimensional color vector. An extreme solution is to binarize the gray levels to obtain black and white images through thresholding.

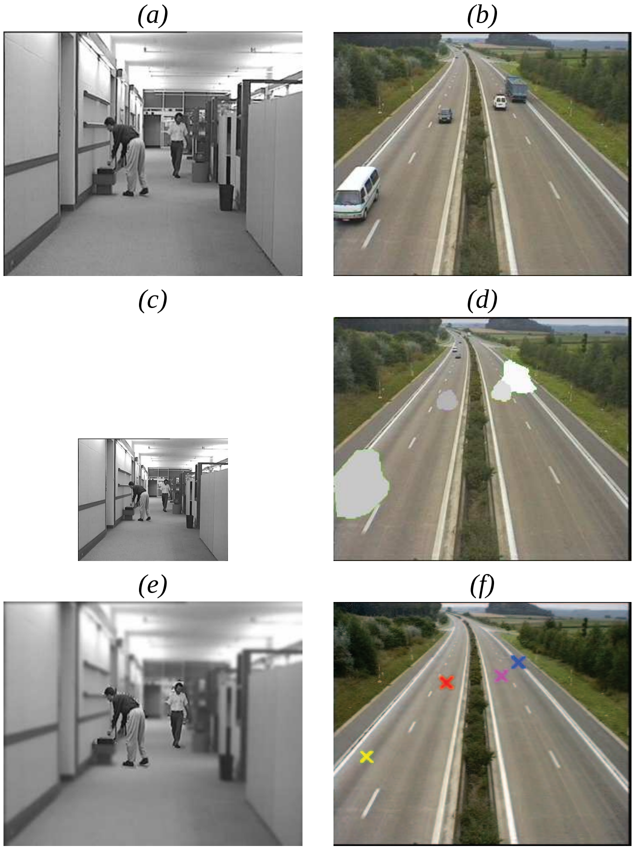

Fig. 2. Transcoding strategies. (a) and (b) Original sequences; (c) Spatial conversion; (d) semantic video objects represented through their shape; (e) semantic video objects superimposed on a low-pass filtered background (f) use of descriptors for transcoding (object identifiers superimposed on the background object)

2.2. Semantic transcoding

The transcoding techniques presented in the previous section convert the media without taking the content into account. On the other hand, semantic transcoding techniques analyze the video content prior to conversion. An example of such analysis is the separation of the video content into two classes of interest, namely foreground and background. Once this separation has been accomplished (as described in Sec. 1), the two classes can be coded differently to better accommodate the way humans perceive visual information, given the available network and device capabilities. The areas belonging to the foreground class, or semantic objects, are used as region of interest. A mask defining the area of interest in the scene based on the semantics of the application (e.g., motion) is first created. Then the areas not included in the mask may either be eliminated, that is set to a constant value, or lowered in importance by using a low-pass filter. The latter solution, simplifies the information in the background while still retaining essential contextual information. An example of this process is shown in Fig. 2(e).

The above mentioned decomposition is used in this paper with an object-based coder as well as with a traditional coder. We will refer to the former case as object-based mode, and to the latter as frame-based mode.

In a *object-based mode*, each semantic video object is encoded separately. The video object corresponding to the background is transmitted to the decoder only once. Then the video objects corresponding to the foreground (moving objects) are transmitted and superimposed on the background object to update the scene. One advantage of this approach is the possibility of controlling the sequencing of objects: the video objects may be encoded with different degree of compression, thus allowing better granularity for the areas in the video that are of more interest to the viewer. Moreover, objects may be decoded in their order of priority, and the relevant content can be viewed without having to reconstruct the entire image (network limitations). Another advantage is the possibility of using a simplified background (appliance limitation), so as to enhance the moving objects. The results of these different approaches in terms of bandwidth requirements are presented in Sec. 3.

In the *frame-based mode*, the decomposition of the scene into meaningful objects helps to support low bandwidth transmission. In our specific implementation, the background is selectively blurred during the encoding process in order to achieve an overall reduction of the required bit rate. The rationale behind this solution is that coding quality can be acceptable if reduced in peripheral regions of interest. This is in line with the behaviour of the human visual system, which assigns greater importance to meaningful areas of a scene.

2.3. Description-based transcoding

Semantic transcoding transforms the input frame-based video into a more suitable format. Such a process is referred to as *intramedia transcoding*. A further processing of the video content may be required before transcoding to cope with limited terminal device or network capabilities. Such processing transforms the foreground objects extracted through semantic segmentation into quantitative descriptors. These quantitative descriptors are transmitted instead of the video content itself. This transformation is referred to *intermedia transcoding*. Intermedia transcoding is the process of converting the media input into another media format. In this specific case, video is transformed into descriptors so as to produce a textual output from the input video. Such textual output can be used not only for transcoding, but also for annotating the video content and for translating the visual content into speech for visually impaired users.

The description of visual content may span different abstraction levels. It can describe the low-level (perceptual) features of the content. These include features such as color, texture, shape and motion. At the other end (i.e., high level of abstraction), it can describe conceptual information of the real-world being captured by the content. Intermediate levels of description can provide models that link low-level features to semantic concepts. In addition, because of the importance of the temporal nature of multimedia and sensitivity of multimedia concepts to context, dynamic aspects of content description also need to be considered.

The process of finding meaningful objects allows us to extract interesting low-level descriptors that summarize the video content and decrease the bandwidth requirements for transmission. In our specific implementation, we use an object identifier and a shape descriptor [6]. The *object identifier* is a unique numerical identifier describing the spatial location of each object in the scene (Fig. 2 (f)). The *shape descriptor* is used to represent the shape of an object, ranging from a bounding box to a polygonal representation with a different number of vertices. A progressive representation is used: the number of vertices corresponding to the best resolution is computed, and any number of vertices smaller that this maximum can be used according to the requirements of the application. An example of result obtained with this process is shown in Fig. 2 (d). In addition to the above, other features such color and texture descriptors may be added in the transcoding process. The choice of these additional features depends on the application at hand.

3. EXPERIMENTAL RESULTS

The experimental results of tests of the proposed semantic video transcoding algorithm (Fig. 1) with the modalities described in Sec. 2 are commented in this section. The ten transcoding strategies used in the tests are summarized in Table 1.

The MPEG-4 test sequence *Hall Monitor* and the MPEG-7 test sequence *Highway* are used to produce the results shown in Fig. 3. Here the bandwidth requirements for the 10 methods under test for 4 seconds are compared.

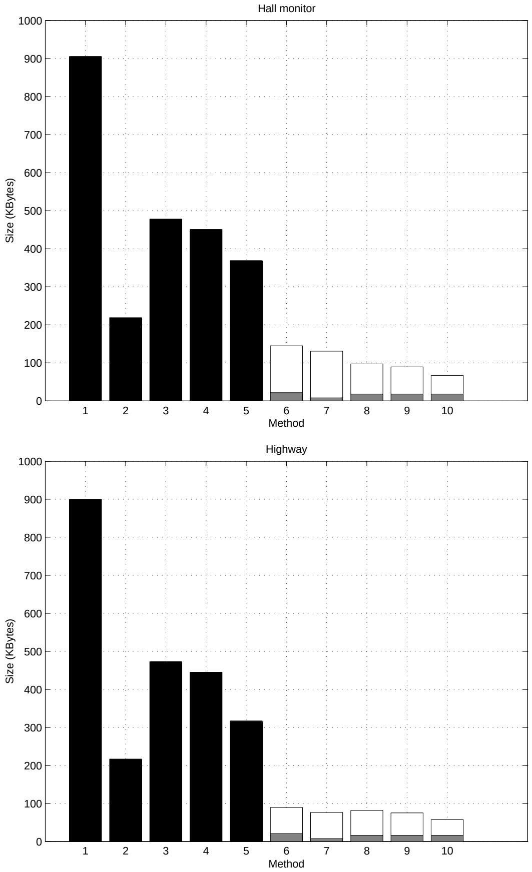

Fig. 3. Bandwidth requirements for the ten transcoding strategies presented in Table 1. Top: Test sequence *Hall monitor*. Bottom: Test sequence *Highway*

The original sequences in CIF format at 25 Hz are encoded with MPEG-1 (1). This represents a reference to compare with the different transcoding solutions. Methods (2) and (3) are content-blind transcoding techniques. Semantic segmentation is used in methods (4) and (5) that still use a frame-based coder (MPEG-1). Semantic segmentation and an object-based coder (MPEG-4) are employed for methods (6) and (7). Finally, the combination of a JPEG2000 encoded background and MPEG-7 BiM encoded descriptors are used in the methods (8) to (10). It is possible to notice a decreasing in the size of the encoded video corresponding to the introduction of semantic segmentation. In particular, methods (4) and (5) use a frame-

method	*description*	*coder*	*frame*	*frame*
			size	*rate* (Hz)
(1)	original	MPEG-1	CIF	25
(2)	spatial conversion	MPEG-1	QCIF	25
(3)	temporal conversion	MPEG-1	CIF	12.5
(4)	semantic transcoding in frame-based mode with fixed background composite before encoding	MPEG-1	CIF	25
(5)	semantic transcoding in frame-based mode with blurred and fixed background composite before encoding	MPEG-1	CIF	25
(6)	semantic transcoding in object-based mode	MPEG-4	CIF	25
(7)	semantic transcoding in object-based mode with blurred background	MPEG-4	CIF	25
(8)	description-based transcoding (20-sided polygons)	MPEG-7	CIF	25
(9)	description-based transcoding (bounding boxes)	MPEG-7	CIF	25
(10)	description-based transcoding (object identifiers)	MPEG-7	CIF	25

Table 1. Summary of the different transcoding strategies. Note: for method (8), (9), and (10) the background is encoded with JPEG2000

based encoder as method (1), but their bandwidth requirement is divided by 2, without the need of changing frame size or frame rate - as in method (2) and (3), respectively. An example of method (5) is reported in Fig. 2(e). For methods (6) and (7) the visualization of the total size is split into two contributions: the background (gray area) and the video objects (white area). Similarly for methods (8)–(10) the white area corresponds to the coding of the descriptors. In the coding process the following tools have been used used: TMPGEnc 2.5, V. 2.59 for MPEG-1 [1], Momusys-FDIS-V1.0 for MPEG-4 [2], MPEG-7 BiM encoder V.02/05/02, and JJ2000 V4.1 for JPEG2000 [3]. The above encoded sequences generated using the proposed transcoder can be found at http://ltswww.epfl.ch/∼andrea/transcoding.html

4. CONCLUSIONS

We proposed a semantic transcoder for adapting an incoming video to various bandwidths and terminal characteristics. In addition to this, we analyzed the impact of semantic segmentation in video transcoding and showed that semantic segmentation is beneficial not only for an object-based encoder but also for a frame-based encoder. The proposed semantic transcoding method is based on semantic segmentation and extracts low-level descriptors from the video content. This makes the algorithm useful not only for video trascoding, but also for video indexing. Furthermore the proposed algorithm can be integrated in a larger framework as a key component of a multimedia documents transcoding system.

Future works include the study and definition of a perceptual metric which accounts for user satisfaction. This metric will be used to automatically select the best transcoding technique by taking into account user preferences, device and network capabilities.

5. REFERENCES

[1] P. van Beek, J. Smith, T. Ebrahimi, T. Suzuki, J. Askelof, “Metadata-driven multimedia access” to appear in *IEEE Signal Processing Magazine*, March 2003.

[2] R. Mohan, J. Smith, C.–S. Li, “Adapting Multimedia Internet Content for Universal Access” *IEEE Transactions on Multimedia*, Vol. 1, N. 1, pp. 104–114, 1999.

[3] A. Vetro, H. Sun, Y. Wang, “Object-Based Transcoding for Adaptable Video Content Delivery” *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 11, N. 3, pp. 387–401, 2001.

[4] A. Cavallaro, O. Steiger, T. Ebrahimi, “Multiple object tracking in complex scenes” *Proceedings of ACM Multimedia*, Juan–Les–Pins (France), pp. 523–532, December 2002.

[5] R. Cucchiara, C. Grana, A. Prati, “Semantic Transcoding for Live Video Server” *Proceedings of ACM Multimedia*, Juan–Les–Pins (France), pp. 223–226, December 2002.

[6] O. Steiger, A. Cavallaro, T. Ebrahimi, “MPEG-7 Description of Generic Video Objects for Scene Reconstruction“ *Proceedings of SPIE Electronic Imaging*, San Jose, California, USA, January 2002.

[1] http://www.tmpgenc.net

[2] http://www.iso.ch/ittf

[3] http://jj2000.epfl.ch