TRANSCODING OF MPEG-2 VIDEO SIGNALS INTO MPEG-4 VISUAL OBJECTS FOR E-LEARNING APPLICATIONS

Nuno P. Santos, Pedro A. Amado Assunção
Institute of Telecommunications / Polytechnic Institute of Leiria,
Pole II, DEEC, Pinhal de Marrocos
3030–290 Coimbra, Portugal
{nunos, amado}@co.it.pt

ABSTRACT

In this paper, we propose a transcoding scheme for e-learning applications, where MPEG-2 video signals are converted into visual objects and then encoded as such by using MPEG-4. The proposed scheme exploits the specific characteristics of the scene contents in order to achieve higher coding efficiency and object manipulation functionality. A spatio-temporal segmentation mechanism is used for extracting two visual objects with semantic meaning in e-learning context. The segmentation algorithm operates in the MPEG-2 compressed domain to produce a binary mask which defines the MPEG-4 visual objects. Then we set appropriate MPEG-4 coding parameters that take into account the specific characteristics of each visual object. The results show a significant improvement in coding efficiency besides the better flexibility provided by the MPEG-4 tools.

1. INTRODUCTION

Nowadays both the MPEG-2 and MPEG-4 standards are widely used in diverse applications. While the former owes its great popularity to the widespread use in Digital Video Broadcasting (DVB) and Digital Versatile Disk – Video (DVD-Video), the latter finds its core applications in multimedia streaming over the internet, mobile networks, as well as in consumer equipment such as video cameras [1,2]. Currently, MPEG-2 technology is also available in the consumer market at affordable costs, which leads to an increasing number of multimedia applications with social and economic relevance.

While MPEG-2 deals with video signals without taking into account the semantic meaning of their contents, MPEG-4 allows coding and processing of visual objects with semantic meaning and arbitrary shapes. Nevertheless, widespread use of this important capability of MPEG-4 is still dependent on reliable and efficient segmentation and tracking tools for dealing with arbitrary objects, since this is not part of the standard. This is perhaps one of the most limiting factors for real-time acquisition, identification and encoding of semantic visual objects extracted from natural video. Therefore, for those application domains that deal with semantic objects, it is necessary to have suitable processing and manipulation tools.

Distance education and e-learning software tools are increasingly important application domains where multimedia technology plays a relevant role. The specific characteristics of such environments and the need for matching different technologies give rise to new types of heterogeneous transcoding and media adaptation schemes for efficient representation and manipulation [3]. This is particularly relevant in most practical applications, where the original scenes are real-time encoded as video frames but separate video objects are necessary for achieving increased flexibility.

The context of this work lies in above mentioned scenari, dealing with transcoding of MPEG-2 coded video into MPEG-4 visual objects for e-learning applications. In this case, the video scenes to be encoded are greatly constrained by the application context. These scenes are mainly characterised by a background (the whiteboard) and a foreground (the teacher). We exploit the fact that both the background and the foreground might be encoded as independent visual objects. Therefore, by taking advantage from the coding tools provided by MPEG-4, we propose a transcoding mechanism for matching MPEG-2 coded signals into MPEG-4 visual objects. A significantly different approach has been recently proposed in [4]. The transcoding scheme relies on a compressed domain spatio-temporal segmentation algorithm. The two visual objects referred to above are extracted from the video frames and then independently encoded by using different coding parameters, according to their inherent characteristics. The results show a good performance taking into account both objective and subjective quality as well as coding efficiency. The target applications range from interactive e-learning multimedia to wireless adaptation through bandwidth and decoding complexity reduction.

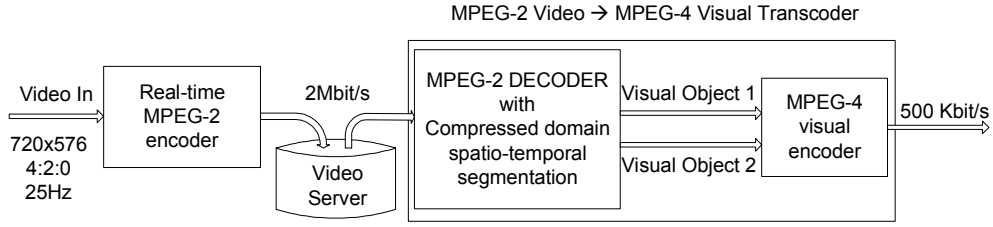

Figure 1: Global system architecture – application context

2. SYSTEM OVERVIEW

The block diagram of the global system is shown in figure 1. The original video is stored in a video server in MPEG-2 format. The transcoder from MPEG-2 video into MPEG-4 visual objects is based on MPEG-2 decoding with segmentation and MPEG-4 encoding. The spatio-temporal segmentation algorithm is implemented in the coded domain taking advantage from the decoder side of the transcoder. The DCT coefficients and motion information is used directly by the segmentation algorithm without further decoding. A similar approach was used in [5]. Then, the segmentation masks are used for defining each visual object which in turn are MPEG-4[1] encoded at a lower rate. Apart from the fact that MPEG-4 is more efficient than MPEG-2, we also take advantage from the fact that the two video objects have much different motion activity. This is described in the next section.

3. THE VISUAL CONTENTS

In the case of e-learning applications, such as those addressed in this work, the video signal is constrained to a particular type of visual information. This consists of a teacher speaking, moving and writing on a whiteboard. From a semantic point of view the teacher is one visual object (foreground) while the whiteboard is another visual object (background). Figure 2 shows one picture of the video sequence that we have used in the experiments. From this figure one can easily identify the two semantic objects of interest: the teacher (foreground) and the whiteboard (background). The following sections describe the main characteristics of these objects and their influence on the choice of appropriate coding parameters.

3.1. The foreground region

The foreground comprises the teacher speaking and writing on the whiteboard. In regard to the subjective quality of this visual object, it should be stressed that motion smoothness is more important than texture accuracy. This means that temporal and spatial quality have different requirements, which can be used for achieving a good compromise between these two parameters. Hence, this object is encoded at full temporal rate (25Hz) in order to produce smooth motion during the most active periods of the scene. These correspond to different types of motion, such as walking, writing on the whiteboard, gesture and speaking.

3.2. The background region

The background is a classic whiteboard where the teacher writes down pedagogical contents for supporting and complementing the oral explanations. The main characteristic of this video object is its relatively slow motion and high texture detail. The slow motion results from the human writing speed on such type of board whereas the high spatial detail is a consequence its specific visual contents, *i.e.*, characters and diagrams written with a marker. Therefore this object can be efficiently encoded by reducing the original temporal rates such that more bits are allocated to encode the texture information, i.e., higher spatial quality.

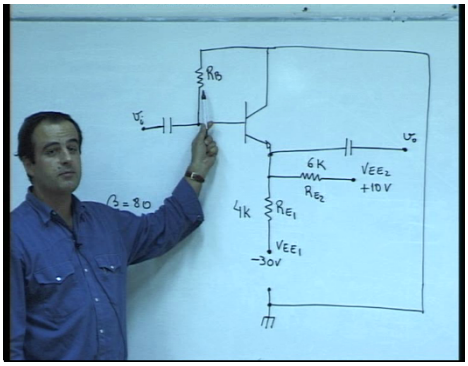

Figure 2: A typical image from a video sequence used in e-learning environment.

[1] We have used a FutureTel hardware encoder for MPEG-2 and MoMuSys software encoder for MPEG-4.

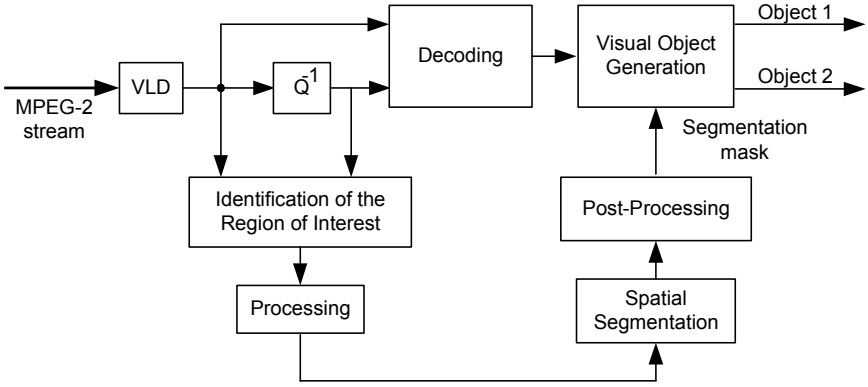

Figure 3: Transcoding through segmentation

4. TRANSCODING THROUGH SEGMENTATION

As mentioned before, the object extraction through segmentation in the transcoder is entirely implemented in the MPEG-2 coded domain. A functional diagram of the algorithm is shown in figure 3 as a sequence of five main operations: 1) partial decoding of the MPEG-2 stream; 2) Identification of the region of interest; 3) processing the region of interest; 4) spatial segmentation and 5) post-processing.

The region of interest is found by analysing a number of Groups of Pictures (GOP) and then finding a coarse area within the pictures where the foreground region is located. This process is based on a temporal segmentation algorithm which identifies a moving region by using the DC images within a predefined temporal window. Since some of the macroblocks identified through this process might not correspond to the moving area because of some false detections, a further refinement processing step is required. This is based on macroblock boundary analysis and predefined semantic rules that are used for checking the relevant macroblocks.

After having the region of interest identified, the next step is basically a spatial segmentation algorithm which labels different regions within the pictures of the MPEG-2 stream by merging in the same region those macroblocks that have similar DC coefficients. This is done within the region of interest previously identified. The post-processing step that occurs afterwards is similar to the previous one.

Then the output of the segmentation algorithm is a segmentation mask which is used for extracting the visual objects from the video signal. Overall this process is similar to that described in [6] though the latter operates in the pixel domain.

5. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed transcoding scheme, we have used MPEG-2 video streams that were previously real-time encoded and stored in a video server. These streams are currently being used for e-learning purposes within the intranet of our campus.

We have carried out several subjective tests in order to find out that, for our hardware encoder, the minimum bandwidth that achieves a good subjective quality is about 2 Mbps. Note that in this type of application poor picture quality is not acceptable because it may lead to additional learning difficulties.

Then the MPEG-2 coded signal was transcoded into an MPEG-4 visual stream combining two objects. The objective is two-fold, *i.e.*, to enable individual object coding and manipulation as well as to increase the scene coding efficiency. While the former is achieved by proper segmentation, the latter greatly depends on both the efficiency of the coding algorithm and the set of coding parameters. In the following sections we present the results obtained from the experiments.

5.1. Segmentation

The spatio-temporal segmentation algorithm was used to produce the segmentation mask for extracting the two objects of interest from the MPEG-2 video stream. These are shown in figures 4 and 5. Note that the mask precision is limited to macroblock level because this is the processing data unit of the algorithm.

5.2. Transcoding efficiency

In order to evaluate the picture quality under a significant transcoding ratio, we have set the output bit rate to 500

Kbps and we compare three different transcoding schemes.

Figure 4. Segmentation mask

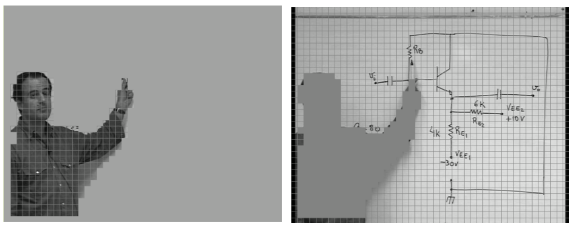

Figure 5. The two visual objects

In all cases we have used the same input video sequence, originally encoded at 2 Mbps. Figure 6 shows the PSNR obtained from the experiments.

Straightforward transcoding from MPEG-2 into MPEG-2 and from MPEG-2 into MPEG-4 using a single rectangular object was used for reference and comparison. In the case of the proposed scheme, we have taken in consideration the inherent characteristics of each visual object, as pointed out in section 3. Then for each object we have set the same output bit rate of 250 kbps but different temporal rates. The foreground was encoded at 25Hz whereas the background was encoded at 6.25Hz. Note that the background area is significantly less that that of foreground. For comparison with the video frames, after decoding the two objects these were combined to form frames again. As we can from figure 6, the proposed scheme achieves a good performance comparing to both references. The PSNR peak in frame 40 is due to no motion activity in the scene during a short period of time which leads to very high picture quality.

6. CONCLUSION

We have proposed a transcoding scheme for e-learning applications which converts MPEG-2 video signals into MPEG-4 visual objects. The experimental results show that a good performance is achieved by choosing different temporal rates for the visual objects according to the specific characteristics of this type of visual scene.

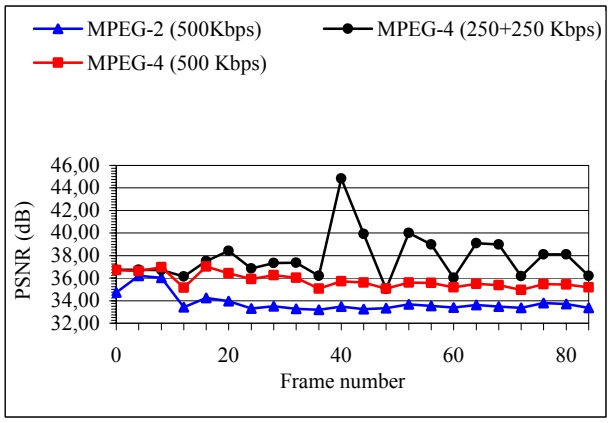

Figure 6: Objective picture quality

On the one hand, by using the proposed transcoding scheme a great deal of reduction in storage capacity can be achieved. On the other hand, for interworking with either networks or user terminals with limited resources, *e.g.*, wireless networks, only the background object along with the audio can transmitted at much lower rates and still delivering an acceptable quality of service.

7. REFERENCES

[1] B. G. Haskell, A. Puri and A. Netravali, *Digital Video-An Introduction to MPEG-2*, Chapman & Hall, New York, 1997.

[2] F. Pereira and T. Ebrahim, *The MPEG-4 Book*, Prentice Hall IMSC Press, New Jersey, 2002.

[3] C. Dorai, V. Oria and V. Neelavalli "Structuralizing Educational Videos Based on Presentation Content", *IEEE International Conference on Image Processing*, Barcelona-Spain, September 2003.

[4] R. Xie, J. Liu and X. Wang, Efficient MPEG-2 to MPEG-4 Compressed Video Transcoding, Visual Communications and Image Processing, Proceedings of SPIE 4671, pp. 192-201, Lugano-Switzerland, July 2003.

[5] X.-D. Yu, L.-Y. Duan, Q. Tian, Robust Moving Video Object Segmentation in the MPEG Compressed Domain, *IEEE International Conference on Image Processing*, Barcelona-Spain, September 2003

[6] M. Kim, J. G. Choi, D. Kim, H. Lee, M. H. Lee, C. Ahn, and Y.-S. Ho, A VOP Generation Tool: Automatic Segmentation of Moving Objects in Image Sequences Based on Spatio-Temporal Information, *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1216-1226, Vol. 9, No 8, December 1999.