A CONTRIBUTION TO VIDEO CODING USING LAYERS*

L. Torres, M. Lecha

Polytechnic University of Catalonia
Department of Signal Theory and Communications
08034 Barcelona, Spain
e.mail: luis@gps.tsc.upc.es

ABSTRACT

Layered representations of video sequences, originally introduced in [1], are important in several applications. Specifically, they are very appropriate for object tracking, object manipulation, content-based scalability and video coding which are among the main functionalities of the future standard MPEG-4, [2] [3]. In addition they will be also useful for MPEG-7 in the corresponding image analysis part. A robust representation of moving images based on layers has been presented in [4]. The objective of this paper is to provide the first results on video coding using the segmentation scheme presented in [4]. To that end, the shape information (alpha map) has been coded using a Multigrid Chain Code approach and the layers (the texture) have been coded using a Shape Adaptive DCT scheme that is being considered as a texture coding approach in the current definition of MPEG-4. Results are provided for a variety of compression ratios and different image quality that prove the validity of the approach.

1. INTRODUCTION

An approach to represent the visual world in terms of objects is the representation in layers proposed by Wang and Adelson in [1]. Each layer contains three different maps: 1) the intensity map, 2) the alpha map, which defines the opacity or transparency of the layer at each point and 3) the velocity map which describes how the map should be warped over time. The layered representation is able to extract different layers corresponding to different objects of the sequence. In each layer all the visible information of the corresponding object is accumulated in one single extended frame. This object-based representation of the audio-visual content has been adopted in MPEG-4. The segmentation of the image sequence will not be standardized what will allow in the future continuous improvements of the object extraction task.

In addition, activities related to the definition phase of MPEG-7 (Multimedia Content Description Interface) have already been started [5] and the standard is to be released in the year 2001. The objective of MPEG 7 is to specify standardized descriptions of various types of multimedia information. This description will be associated with the content itself, to allow fast and efficient searching for material that is of interest to the user. As in the case of MPEG-4, the segmentation part will not be included in the standard, but MPEG-7 will need object-based representations of the video content. It is in this context that the layered representation of a video sequence may play an important role as a powerful approach to the segmentation part needed to input segmented objects into the standardized part of MPEG-4 and MPEG-7.

Previous work on the layered representation of a video sequence has dealt almost exclusively with the obtention of the layers [1] [4]. That is, with the motion estimation and segmentation part of the scheme needed to extract the objects and to represent the video sequence in layers. Notice that this representation allows manipulation of the content of the scene and once the layers are formed, each frame is defined only through the motion parameters what gives a very compact representation useful in video coding applications. It is in this context that the objective of this paper is to provide further insight into the layered representation by showing its application to video coding. This application provides a compact framework for object-based representation of visual material by giving, in addition to coding, capabilities of object manipulation useful in the context of MPEG-4. For clarity reasons, Figure 1 shows an example of the layered representation of the mobile calendar sequence.

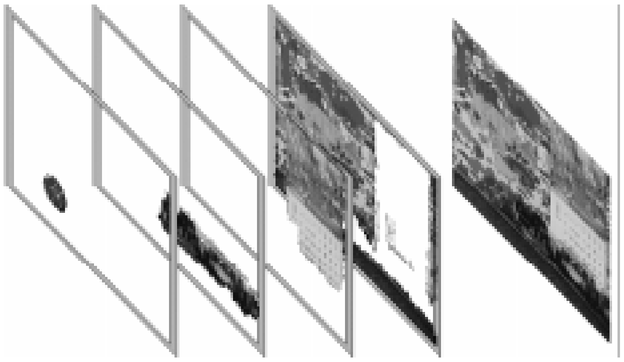

Figure 1. The layered representation of the mobile calendar sequence

The layered representation paves the way for efficient bit assignment schemes based on objects (layers). This is, bit assignment can be done for every object depending on the importance of each object in the scene. This approach provides video content scalability as well as spatial and temporal scalability. As a side comment it is important to remark that the visual quality of the decoded reconstructed image will depend

* *This paper has been supported by Grant TIC95-1022-C05 of the Spanish Government*

enormously on the quality of the segmentation scheme used to generate the layered representation of the image, and, of course, on the number of bits available. This means that in order to obtain a successful video coding scheme using layers, the quality of the segmentation used is of paramount importance.

In the context of the use of layers for video coding, three different information need to be coded. First is the intensity map, second is the alpha map and third is the velocity map. This way of coding a video sequence by using in a separate way the texture, the segmentation (partition) and the motion information is a class of the so-called Second Generation Video Coding Schemes [6]. The objective of this paper is to show that, in addition to object manipulation and tracking, the layer formation scheme presented in [4] exhibits good performance in the context of video coding. To this end, we have applied efficient texture and contour coding techniques to encode the intensity and alpha map of the layer information. In particular, a Shape Adaptive DCT scheme has been considered to encode the texture information, and a Multigrid Chain Code approach for the contour part. The motion information has been coded using uniform quantizers followed by entropy coding.

The paper is divided in the following sections. Section 2 gives a very short description of our approach presented in [4] to provide robust motion estimation and segmentation algorithms to obtain the layered representation of the video sequence. Section 3 provides details on the Shape Adaptive DCT scheme used for texture coding and of the Multigrid Chain Code approach for the contour part along with coding of the motion parameters. Section 4 provides video coding results and conclusions.

2. ROBUST MOTION ESTIMATION AND SEGMENTATION

This section summarizes very briefly our approach related to the generation of layers using efficient techniques of motion estimation and segmentation. For further details, the reader is referred to [6]. Our construction of the layers follows conceptually the approach presented in [1] but using more efficient and robust techniques. The process implies two basic operations: 1) a local motion estimation and 2) a motion segmentation by affine model fitting. In order to estimate the motion field of the scene, we compute first an initial approximation of the motion of some pixels located in a rectangular grid using block matching techniques. The motion vectors of the remaining pixels are interpolated from these first ones. Due to local minima of the cost function defined, some motion vectors may be erroneous. To improve these motion vectors, a motion gradient is found for each pixel. Then for those pixels whose motion vectors are above the mean of all the gradients, the motion vectors are recalculated using a larger neighborhood. The interpolated motion vectors are refined with integer precision. The final motion vector for each pixel of the whole image is obtained through a refinement process based on one third-pixel accuracy.

The next step is to segment the motion vector field by fitting an affine motion model. The objective is that those regions of the image with the same affine model will be grouped. As we do not have a reliable initial image partition, the image is initially partitioned in rectangular regions. For each of these rectangular regions, the best affine model in the minimum square sense is found through an iterative k-means clustering algorithm. Once a set of affine models have been found, similar models are grouped based in a mean-square distance between the motion vectors of the pixels belonging to each of the models being compared. Two models are grouped each time. The ordering in which all the model pairs are grouped is of paramount importance as it will fix the stability and convergence of the technique. We have provided in [4] adequate tools for ordering purposes.

The robust segmentation technique explained above provides a series of non-overlapped regions that will be used to form the final layers. Our layer synthesis approach follows very closely to the original one presented in [1]. Once the layers are formed, they are ready to be encoded. Section 3 provides this explanation.

3. CODING OF THE LAYER INFORMATION

In the context of the use of layers for video coding, three different information need to be coded. First is the intensity map, second is the alpha map and third is the velocity map.

3.1 Coding of the intensity information

In a layered image sequence representation approach, each frame is defined only through the corresponding motion parameters. Thus, the intensity information that needs to be coded is just an intraframe image associated to each layer. If the layer formation process has done a good work, there is no need to encode the prediction error. This is the case for layers that follow very closely the affine motion model defined in our scheme. To encode the intensity information, a texture coding technique that takes into account the arbitrary shape information associated to each layer will have to be used. We have selected the Shape Adaptive DCT (SA-DCT) for our purposes [7]. This technique is also being considered as a texture coding technique in current developments of MPEG-4 standardization activities.

SA-DCT is based on predefined orthogonal sets of DCT basis functions. The algorithm is fast to compute and does not require more computations than a normal DCT. Images are separated into adjacent square blocks and only the full blocks contained within an object or segment region are encoded using a standard DCT. Blocks containing the boundary of segmented regions are encoded separately using SA-DCT. In this case after shifting and alignment of the vectors that compose the object, a unidimensional DCT is applied to both the vertical and horizontal directions. For more details the reader is referred to [7].

3.2 Coding of the shape information

To encode the shape information (alpha map), several options are available. Common practice in shape coding is to use a lossless chain code approach. This approach achieves an average of 1.4 bits per contour pixel and has the advantage of being lossless. We have preferred to use a lossy approach that provides practically lossless visual results. This scheme is known as Multigrid Chain Code (MCC) and has been also considered in

MPEG-4 [8]. MCC uses a specific contour grid such that some movements along the contour grid have larger steps than others. As a result, the number of steps necessary to describe the contour, that is the number of symbols to code, will be low. For further details the reader is referred to [8].

3.3 Coding of the motion information

Motion information has to be transmitted in order to know how the intensity map should be warped over time. The motion parameters can be sent without compression because they only represent six numbers per layer per frame. Entropy coding techniques can be used for further compression.

4. RESULTS AND CONCLUSIONS

To test the validity of the layered approach for video compression, we have encoded different video sequences. In particular, results will be shown for the first 30 frames of the Flower Garden sequence used in MPEG standardization activities. The size of each frame is 720x288 pixels, 8 bits per pixel, which gives about 41.5 Mbits/s for the original image. Our layer formation scheme does a quite good on this sequence as it is mainly composed of translational motion. Four layers have been extracted: Field, Sky, Houses and Trees. Table 1 shows the bits used to encode the texture information of each of the extracted layers and for different quantization steps (QP) of the DCT coefficients. Table 2 shows the results for the alpha map (shape) encoding of each layer using the MCC approach. Motion parameters have not been compressed and are coded with 9.9 Kbits/s. Figure 2 shows the original 15th frame of the Flower Garden sequence. Figures 3 and 4 show the same frame coded at 312 and 179 Kbits/s. No prediction error has been encoded in any of these sequences. Notice the introduction of block effect at this last rate. Some other artifacts are also visible. For instance, the men have not been correctly reconstructed. This is due to errors in the segmentation scheme.

The number of bits allocated to each layer has been uniform. That is, all the layers have been coded with the same quantization step. In some applications, layers may be represented with different quality. If this is the case, the layered representation provides a useful framework in which different number of bits may be assigned to the layers, thus providing object-based bit assignment.

The scheme has been applied to other sequences with more complex motion. If the motion of the sequence is far from the affine model used in the segmentation, then the prediction error between two different frames has to be encoded. Work is in progress to cope with this situation. Some manual user interaction in the segmentation stage proves also useful in some occasions.

As a conclusion, let us say that the proposed representation provides a useful framework for object-based coding. In addition to compression, object tracking and manipulation is also provided by the layered representation what gives to the approach a broad range of useful applications.

Layer	QP = 10	QP = 15	QP = 20	QP = 25	QP = 30
Field	170649	117084	87748	69450	56819
Sky	18877	16989	15845	15279	14583
Houses	68069	46244	35100	28765	24744
Trees	36675	25483	19751	16596	14531
TOTAL	294270	205800	158444	130090	110677

Table 1. Bits used to encode the texture information of each layer as a function of the quantization step

Layers	Bits	Contour pixels	Bits/pixel
Field	2985	2545	1.17
Sky	4990	4326	1.15
Houses	4255	3301	1.29
Trees	3406	2945	1.16
TOTAL	15636	13117	1.19

Table 2. Bits used to encode the shape information of each layer

5. REFERENCES

[1] J. Wang and E. Adelson "Representing moving images with layers". *IEEE Transactions on Image Processing*, 3(5):625 – 638, 1994.

[2] F. Pereira "MPEG 4: A New challenge for the representation of audio-visual information". *Picture Coding Symposium*, Melbourne, Australia, 1996.

[3] ISO/IEC JTC1/SC29/WG11. MPEG Requirements Group. "Proposal package description", July 1995.

[4] L. Torres, D. García and A. Mates "A robust motion estimation and segmentation approach to represent moving images with layers". *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, pages 2981-2984, April 1997.

[5] ISO/IEC JTC1/SC29/WG11. MPEG Requirements Group. "MPEG-7: Context and Objectives", Doc. ISO/MPEG N1733, MPEG Stockholm Meeting, July 1997".

[6] L. Torres and M. Kunt *"Video Coding: The second generation approach*". Kluwer Academic Publishers, 1996.

[7] T. Sikora and B. Makai "Shape adaptive DCT for generic coding of video". *IEEE Transactions on Circuits and Systems for Video Technology*, 5(1):59 – 62, 1995.

[8] P. Salembier, F. Marqués and A. Gasull "Coding of partition sequences" in L. Torres and M. Kunt *Video Coding: The second generation approach*. Kluwer Academic Publishers, 1996.

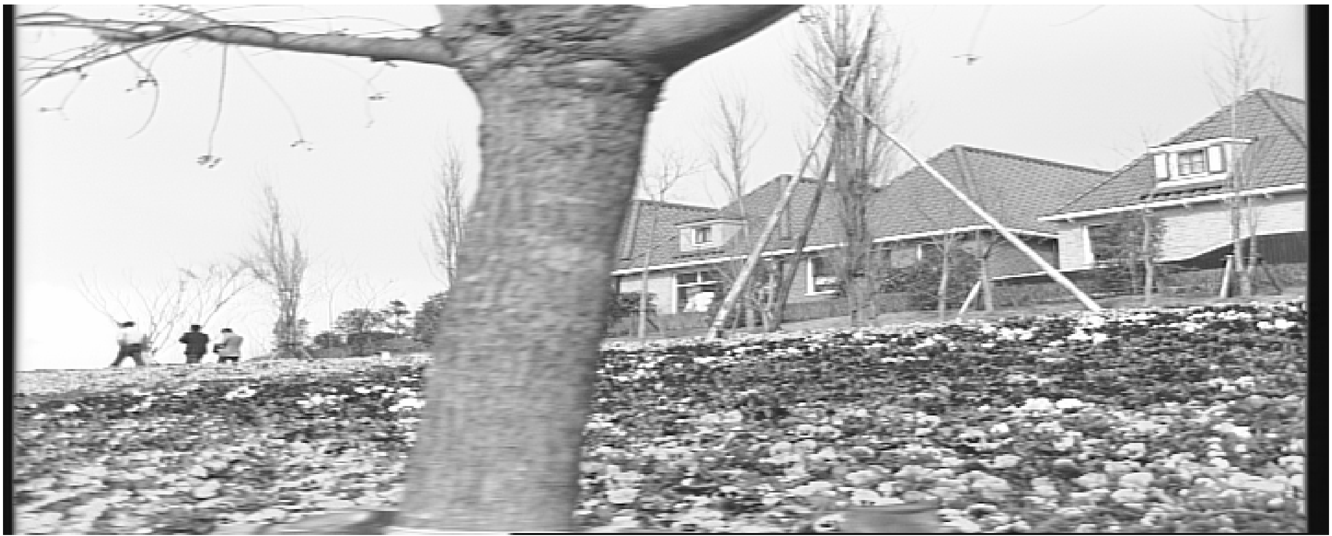

Figure 2. Original 15th frame of the Flower Garden sequence

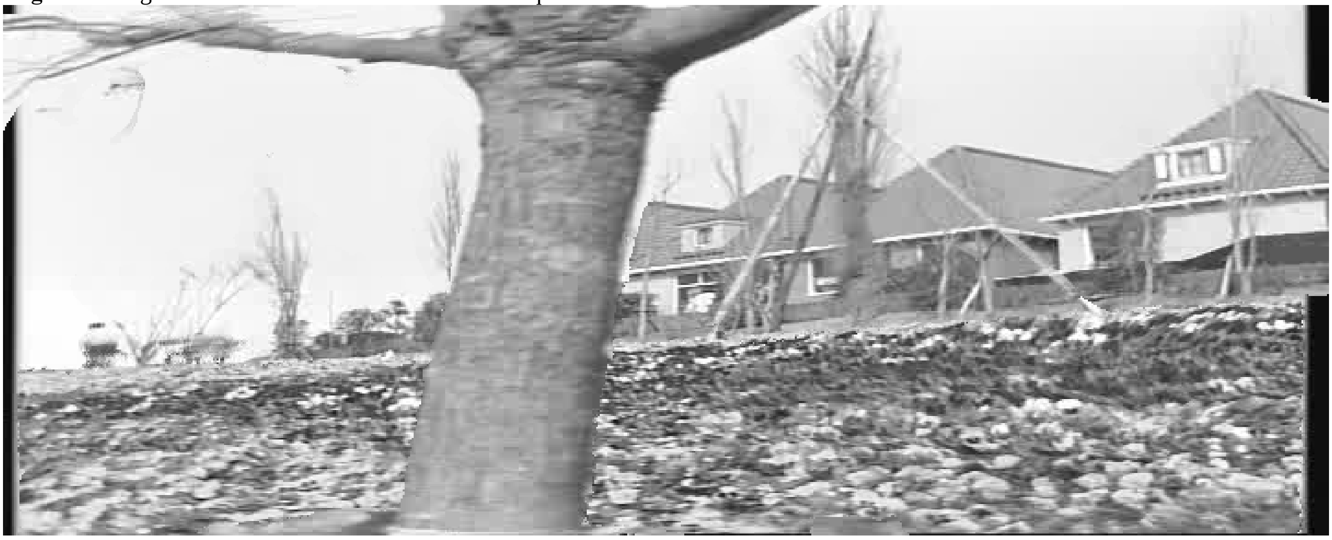

Figure 3. Reconstructed frame at 312 Kbits/s.

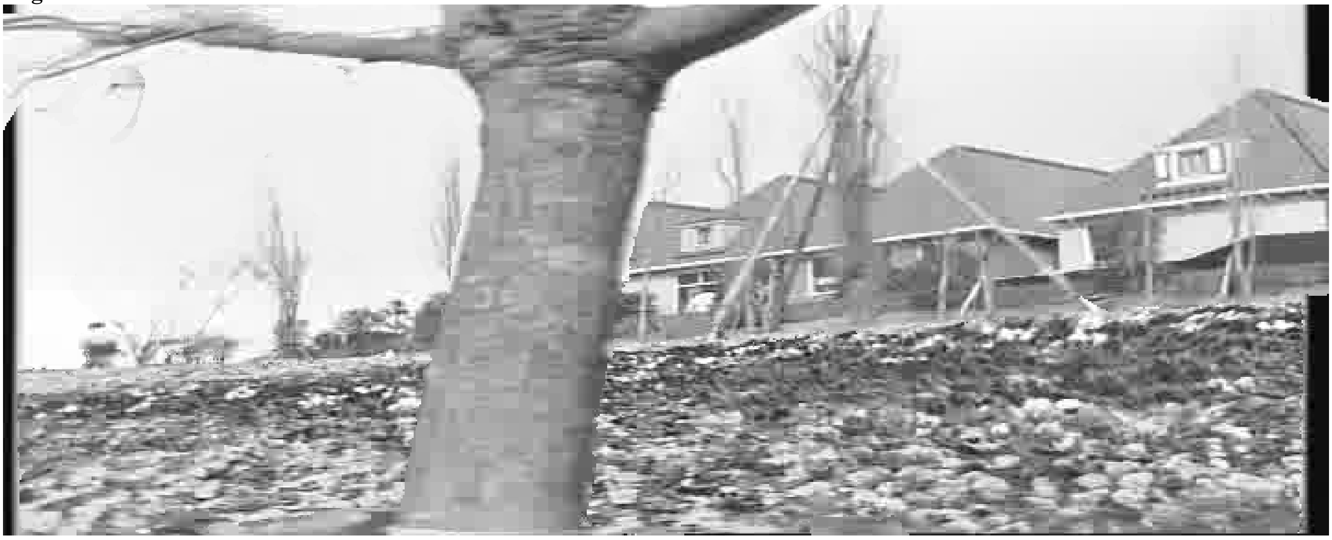

Figure 4. Reconstructed frame at 179 Kbits/s.