TREE BASED MOTION COMPENSATED VIDEO CODING

R. Leonardi, Associate Professor†‡, & H. Chen, MTS‡

†Signals & Communications Lab., Dept. of Electronics for Automation,
University of Brescia, Brescia, I-25123, Italy
Ph: +(39-30) 371-5434 Fax: +(39-30) 380-014
e-mail: leonardi@icil64.cilea.it
‡Visual Communications Research Dept.
AT&T Bell Labs/ Visual Comm. Res. Dept.
Holmdel, NJ 07733, U.S.A.

ABSTRACT

In this paper, we present a novel technique to encode video sequences, that performs a region-based decomposition of each frame on the basis of motion information. Using the segmentation map, any region in a frame to be encoded will be predicted from a single reference frame, using motion compensated prediction. The use of a single reference frame avoids feedback of the prediction error information in the prediction of successive frames. Coding is simply obtained by describing the segmentation map and the associated motion information. Error information will not be provided for low bit-rate applications. The segmentation map is described using a quadtree structure. Within such a tree structure, we show how motion information can be predicted either spatially or temporally, so as to minimize redundancy of information. The motion and segmentation information are estimated on the basis of a two stage process using the frame to be encoded and the reference frame: 1) a hierarchical top-down decomposition; 2) a bottom-up merging strategy. The proposed method is used to encode to encode QCIF video sequences with a reasonable quality at a 10 frame/s rate using roughly 20 kbit/s.

1. INTRODUCTION

There is a variety of applications using video information where the available bandwidth remains very limited, typically 19.2 kbit/s[1]. In this context, algorithms derived from traditional video compression schemes are likely not to perform adequately at these bit-rates.

Efforts have been made to modify CCITT recommendation H.261 for video-telephony so as to work at bit-rates lower than 64 kbit/s [1], by changing the video sequence format and adapting the parameters of the coding scheme. H.261 decoding recommendation is based upon a hybrid motion-compensated predictive strategy with DCT coding of the prediction error. The simulation results of the corresponding SIM3 are still uncertain at this point especially when significant activity is present in the source material. A still troublesome issue is to design a buffer control strategy at these bit-rates that can manage large amounts of new information. Besides, the extent of SIM3 is limited to source material which remains in the video-telephony context.

It is essential to distinguish applications where "real-time" transmission is necessary, like video-telephony, so that a satisfactory interactive communication is feasible, from those that are not necessarily stringent in the time-delay for image reconstruction at the receiver end. Consider for example video browsing for multimedia purposes. In this context, encoding time can be quite significant, and schemes that try to take advantage of temporal correlation over several frames can be considered.

Even if one considers other state of the art video coding algorithms (such as in MPEG1 orMPEG2), that may be used for other classes of video sequences, it is difficult to imagine how these can be modified so as to work at lower bit-rates given the tremendous amount of overhead information they use. These algorithms mainly improve the quantization strategy of the information to be coded by making such quantization adaptive to the frame content. These algorithms

This work was supported by AT&T Bell Laboratories and the Italian Council for National Research.

[1]Even if transmission over twisted pairs seems to be possible with large bandwidth for distances of less than 1 mile (given appropriate adaptive equalization techniques), it is still difficult to provide large bandwidth across traditional switches, and very unlikely to have it for wireless communication.

provide also a better rendering of motion information through the use of motion-compensated interpolation techniques. However, they deal with image sequences in a block-based fashion.

This work will introduce a new approach for video compression, which can be applied for both types of applications. In fact, we present a methodology that can cope with short encoding delays (assuming adequate hardware) while providing acceptable performance at low bit-rate for video compression.

Improving compression for image sequences requires using additional properties than the simple block-based motion information, and underlying stationarity assumption which is commonly used in the more classical schemes. Some attempts have been made to insert a priori knowledge for classes of video sequences, as in the wireframe models [2] or the head-and-shoulder models, used in the analysis-synthesis proposal of SIMOC [3]. We do not want to restrain the approach here to a limited class of video sequences, but use an often forgotten property of video sequences generated from a natural scene, i.e. the structure information. By structure, we propose to identify the spatial arrangement of pixels that exhibit a common property (texture or motion are two examples), so as to obtain an object based decomposition of each frame, i.e. a segmentation of such frame. Motion compensated prediction can then be applied to each region of the segmentation.

If translational motion has been extensively used for block-based motion compensation, more elaborate motion models have not been considered, so that it is difficult to describe deformations of objects between successive frames; to keep a fixed block size has been another limitation for the correct estimation of motion information as objects in an image are rarely squares of fixed size. This has suggested to some to identify objects in a given sequence by appropriate segmentation techniques that rely on a joint motion and luminance analysis. These objects are then tracked over the sequence of frames. This approach allows still for interactive communication, as long as this motion tracking process remains limited to a few frames (typically 3-5 frames) [4]. Given the 3-dimensional structure of the natural scenes from which the frames are generated, it is clear that even though the information is treated in an object based fashion, the motion description is not sufficient to generate from a single frame, any new frame. In effect, there are uncovered areas that represent new information to be transmitted. Transmitting video information requires therefore sending three items:

1. the object shape or *region description* resulting from a segmentation process, on a frame by frame basis. It is not necessary to transmit a complete segmentation for each frame, as 2 successive segmentation maps are in general correlated.
2. the *motion description* of each region.
3. the *new information* such as uncovered background, or error information resulting from an inappropriate motion/segmentation estimation.

Accordingly, we have constructed the following video coding scheme: a) a motion based segmentation algorithm so as to obtain for each frame in the sequence the moving objects and their apparent motion; b) a sophisticated coding procedure to describe the segmentation shape and the motion information associated to each region; c) from the encoded information, a motion-compensated algorithm that allows to reconstruct the video sequence.

The overall coding procedure is identified in the block diagram of Figure 1.

Figure 1: Video coding algorithm

The paper is organized as follows: In the next section, we describe the motion segmentation process using a quadtree decomposition of the frame to be encoded. Section 3 discusses then the coding strategy with a discussion of some quality compression performance trade-offs quoting some simulation results that have been made, while section 4 provides a summary and a direction for future research.

2. SEGMENTATION STRATEGY

2.1. A unique segmentation map

The segmentation process is necessary so as to base the motion compensation process on a region basis, rather than on a macro block (16×16 block) one.

If this idea is simple in its concept, there are some very delicate issues to be addressed. In particular,

sending information according to items 1, 2 and 3 in the previous section may require transmitting 3 different segmentations: The region description one, the motion description one and the error description one. These may differ significantly from each other unless specific constraints are imposed. From traditional segmentation based image coding models [5], the description of the segmentation map is very expensive, typically about 1 bit per region boundary point. It seems therefore necessary to reduce the number of such maps to be encoded.

To guarantee a unique segmentation map, we suggest to

- base the segmentation solely on motion information. Only when the apparent motion between 2 or more successive frames is insufficient to define accurate region boundaries, luminance information of a single frame may be used. However, this is not necessary as an inaccurate region boundary obtained on the basis of motion information does not mean a poor reconstruction of the frame after region-based motion compensation.

- compute the segmentation for the frame to be encoded using the past or future frames as references, rather than predict the current frame by displacing regions of the future or past frames. This allows any error information after compensation to be limited within the region boundaries of the motion segmentation. An on/off flag is then sufficient to describe whether or not such error information needs to be coded for any given region of the original segmentation.

2.2. Quadtree based segmentation

To limit the complexity of the segmentation algorithm, and to have a structured data representation, that may further benefit from a better correlation between segmentation maps of successive frames, we embed the segmentation graph in a tree structure. This seems further attractive, as certain tree structures provide a good trade-off between region and luminance description [6] (in our case, it may provide a good trade-off between region and motion description). For simplicity, we select quadtrees (but more general structures could have been considered such as BSP trees), and have constructed a quadtree motion segmentation algorithm, that works in a top-down fashion followed by a bottom-up one.

The top-down part of the algorithm works as following: For any given frame, a multiscale block based motion estimation is iteratively computed with respect to a previous reference frame. Starting from the entire current frame, blocks are optimally motion-compensated using the reference frame[2]: Whenever the block minimum motion-compensated error energy falls above a dynamic threshold, the block is split into 4 sub-blocks, and the procedure is then iterated. The dynamic threshold is varied according to the block dimension, so as to allow larger error energies for smaller block sizes. The splitting process can be terminated when the block size reaches a lower bound (typically 2×2.) The adaptation of the threshold is essential to obtain a good trade-off between noisy estimation of small size blocks and detection of small moving objects in large size blocks. At the end of this process, a quadtree is formed, with a motion description (i.e. a displacement vector) and a block-based prediction error assigned to each leaf node.

The quadtree is then scanned in a bottom-up fashion so as to successively merge 4 children into their parent node, whenever the motion information assigned to the 4 children of a given parent node are very similar. As a result a quadtree segmentation has been reached, each leaf node in the tree defining a square region. To each leaf node, one can assign both a displacement vector information and a prediction error signal.

2.2.1. Top-down decomposition

For a $2^p \times 2^p$ frame size, by calling $Q_{i,k,l}$ the tree node of level i (level 0 is the root node of the quadtree) and position j, k defined by the coordinates of the top left point of each block, its children are defined by the tree nodes $Q_{i+1,k,l}, Q_{i+1,k+2^{p-i},l}, Q_{i+1,k,l+2^{p-i}}$, and $Q_{i+1,k+2^{p-i},l+2^{p-i}}$.

In the top-down approach, an optimum displacement vector is estimated for each $2^{p-i} \times 2^{p-i}$ blocks corresponding to all tree nodes to be processed at a given level i. Let us call $E_{i,j,k}$ and $\mathbf{d}_{i,j,k}$, the minimum displacement difference energy and associated displacement vector of one such block at level i with respect to the reference frame, respectively. The motion estimation is iterated at level $i+1$ if $E_{i,j,k} > T_i = 2^{p-i} \times T_{i-1}$.

2.2.2. Bottom-up merging

In the bottom-up part of the algorithm, children nodes $Q_{i+1,k,l}, Q_{i+1,k+2^{p-i},l}, Q_{i+1,k,l+2^{p-i}}$, and $Q_{i+1,k+2^{p-i},l+2^{p-i}}$ are merged into their parent node $Q_{i,k,l}$ if $\mathbf{d}_{i+1,k,l}, \mathbf{d}_{i+1,k+2^{p-i},l}, \mathbf{d}_{i+1,k,l+2^{p-i}}$, and $\mathbf{d}_{i+1,k+2^{p-i},l+2^{p-i}}$ are within some small fraction from each other (typically ±1 for each component). In case of merging, $\mathbf{d}_{i,k,l}$ corresponds to the integer rounded value of the four children displacement vector average.

[2]Optimality is guaranteed in the sense that it minimizes the displacement error energy of a same size block from the reference frame within a certain range of possible displacements.

Given the randomness of the displacement estimates for small size blocks (large values of i), a better merging strategy is under investigation. This is obtained by selecting for each set of 4 children of a given parent, one of the 4 displacement vectors found for each one of them and use it to displace by the corresponding amount their parent node. If the new prediction error energy does not increase significantly, the 4 children are merged into one. The problem lies in selecting in an automatic fashion the maximum diversion of the prediction error with respect to the optimal one.

For small size blocks, the motion estimate may not relate at all to a physical displacement. In fact, small blocks may correspond to uncovered areas in the current frame. The displacement vector allows then by compensation to place the best match of a same size block within a certain spatial range from the reference frame. The set of all possible blocks of same size from the reference frame can be seen as a locally adapted dictionary of a vector quantizer, the displaced blocks being codewords of the dictionary. For this reason, we assume that even though uncovered areas of the current frame are not present in the reference frame, there still are codewords of the dictionary that are good matches for small size blocks. This way, it is unlikely that the non-transmission of error information will lead to unacceptable errors.

3. CODING STRATEGY

A multiscale spatio-temporal coding algorithm has been designed to describe the motion information jointly with the quadtree segmentation map. While describing the quadtree structure, a spatio-temporal prediction of the motion information is performed with very little overhead.

In the event there is enough bandwidth to encode some error information, we suggest to define the segmentation error map on the basis of all contiguous squares exhibiting the same displacement value. Each relabelled region for which the error signal is too high can be encoded in a region based fashion, using a set of basis functions that can be made orthonormal with respect to the support of the region [7].

3.1. Spatio-temporal prediction of motion information

To encode the quadtree segmentation map with associated motion information, the tree is traversed in a top-down fashion. At each level, all nodes are successively analyzed, so as to specify whether they have descendants or they define leaf nodes. In the former case, a '0' bit is transmitted whereas in the latter a '1' is transmitted. If a '0' is transmitted, the 4 children of the current node will be processed when the next level of the tree is reached. If a '1' bit has been transmitted, the motion information assigned to the corresponding leaf node is encoded. For this purpose, we use the spatio-temporal correlation of motion information, a 3 symbol alphabet is defined: 'T', 't', and 's'.

- If 'T' has been transmitted with a '0' bit, the motion vector can be predicted at no additional cost as the average of the previously encoded frame motion vectors that have been temporally extrapolated to the current frame and fall within the block corresponding to the leaf node being currently encoded.

- If 't' has been transmitted with a '$1-0$' 2-bit word, the motion vector is still temporally estimated, but coding of the difference value with respect to the average temporal estimate is further encoded component-wise. This selection is made with respect to the 's' selection only when the spatial prediction is less accurate than the temporal one.

- If 's' has been transmitted with a '$1-1$' 2-bit word, the displacement is spatially predicted from the displacement of the largest leaf node which was contiguous in space to the current node. This selection may not be the optimal one, but it is reasonable and can be used by the decoder at no additional cost, as a spatial prediction may result more accurate from a larger block than a smaller one. The difference value with respect to the current block displacement value is further encoded component-wise.

The transmission of a 'T' or 't' symbol is considered only for blocks larger than a certain size (typically the smallest considered size), and is used to handle the constant velocity of most objects over time. As small size blocks may have unreliable motion estimates, it is unlikely that a temporal prediction may work in this case. In such a case, only a spatial prediction is made, with no need therefore to transmit the 'T', 't' or 's' overhead.

3.2. Problems of the reference frame

It is essential to note that the algorithm as suggested may give satisfactory performance if the reference frame is of high quality, which means that it has been sent at the beginning of the transmission. If a scene cut occurs in the video sequence, a new reference frame has to be sent, to provide the future basis for prediction.

This will result for a small channel capacity in a significant delay (about 0.5 s) before video information with motion can be recovered at the receiver end.

It is also necessary to note that after a certain time, the original reference frame may be too far apart from the current frame being encoded, so that a portion of the available bandwidth should be used to update the reference frame. The bit-allocation to split the available channel capacity in 2 layers of transmission, one for the reference frame update, the other for the residual coding of the current frame is a very complex issue, currently under investigation.

3.3. Simulation results

Using few high quality reference frames (one every 2 seconds), QCIF video-telephony sequences have been encoded at a rate of 10 frames/s and a bit-rate of about 20 kbit/s with a satisfactory quality (The cost for encoding the high quality reference frames has been neglected at this point).

4. CONCLUSION

We have described a quadtree based motion predictive scheme to encode video sequences at low bit-rates. The novelty of the approach lies both in the joint motion estimation and segmentation procedure, and in the joint coding of the motion information and segmentation map. We have suggested to predict spatially or temporally the motion information for each square region in the quadtree using neighboring size blocks or the average value of motion information extrapolated from previously encoded frames. A rigorous strategy is adopted to obtain the proper motion estimate at no additional cost in terms of overhead.

Extensions of this encoding scheme are considered so as to include error information and to design a sophisticated buffer control strategy to accommodate increased complexity of the video sequence (by redefining the high quality reference frame). The use of time correlation is being investigated as well to predict the change of the segmentation map (the quadtree decomposition) over time.

References

[1] *Simulation model for very low bit-rate image coding (SIM3)*, CCITT SGXV WPXV/1: Special Rapporteur Group for Very Low Bit-rate Visual Telephony.

[2] K. Aizawa, "Model Based Coding", to be published in the *Handbook of Visual Communications*, Ed.s H.-M. Hang and J. Woods.

[3] J. Ostermann, "SIMOC: A European Initiative towards MPEG-4 by COST 211", ISO/IEC JTC1/SC29/WG11, MPEG-4 seminar, Paris, Mar. 1994.

[4] P. Brigger and M. Kunt, "Contour Image Sequence Coding Using the Geodesic Morphological Skeleton", in *Proc. of the VLBV94 workshop, University of Essex, Colchester (UK)*, paper 3.1, Apr. 1994.

[5] R. Leonardi, M. Eden, and M. Kocher: *Coding a Contour Graph with No Address Assignments*, AT&T Bell Laboratories Technical Memorandum, Doc. No. 11355-901115-12TM, Nov. 1990.

[6] H. Radha, R. Leonardi, and M. Vetterli: *Coding Images Using BSP Trees*, to be submitted to IEEE Transactions on Image Processing.

[7] M. Gilge, T. Engelhardt, and R. Mehlan, "Coding of arbitrarily shaped image segments based on a generalized orthogonal transform", *Signal Processing: Image Communication*, 1(2): 153-180, Oct. 1989.