Temporal Prediction of Video Sequences Using an Image Warping Technique Based on Color Segmentation

N. Herodotou, and A.N. Venetsanopoulos
Digital Signal & Image Processing Laboratory
Department of Electrical and Computer Engineering
University of Toronto
10 King's College Road
Toronto, Ontario
M5S 3G4
CANADA
E-mail: nicos@dsp.toronto.edu
E-mail: anv@dsp.toronto.edu
URL: http://www.comm.toronto.edu/ ~ dsp/dsp.html

Abstract. An image warping technique based on segmented regions is introduced for the temporal prediction of videophone-type sequences. At the encoder, a set of control points are determined from the previous frame and their corresponding best matched points are determined from the current frame. The selection process of these points is achieved by segmenting the previous frame into different regions using a color segmentation technique based on a recursive histogramming approach and the control points are subsequently chosen along the region boundaries. The spatial offset of these points between the previous and current frame are represented as motion vectors. The facial area is intraframe encoded to avoid distortions of the facial features which convey critical information. At the decoder, the same segmentation and control point selection algorithm is used along with the motion vectors in order to find the region boundaries of the predicted frame. The facial area is decoded and an affine transformation is finally used to determine the remaining regions and form the predicted frame. This technique produces results that are free from blocking artifacts as in the conventional block matching method, with only a moderate increase in computational complexity.

1 Introduction

A number of multimedia applications have newly emerged due to the recent advances in the area of mobile communications and the tremendous growth of the *Internet*. Applications such as portable communicators, video email, and video databases, to name a few, have placed even greater demands for more effective video coding schemes. However, future coding techniques must focus on providing better ways to represent and exchange visual information in addition to efficient compression methods. These efforts aim to provide the user with greater flexibility for *content-based* access and manipulation of multimedia data as in the proposed MPEG 4 and future MPEG 7 standards [1].

In conventional coding methods such as H.261 and MPEG 1 and 2, video sequences are compressed using an information-theoretic-based approach, that is, by exploiting the stochastic properties of the signals. Recently, however, greater attention has been paid to a newer generation of coding schemes which are *object-based* [2, 3]. These methods rely on the techniques of image analysis and computer graphics to represent the image signals using their structural features such as contours and regions. In this latter approach, an input video sequence is first segmented into an appropriate set of arbitrarily shaped regions. The features of each region such as shape, motion, and texture/color information are subsequently used in the encoding process. Thus, the success of an object-based method depends largely on the segmentation of the scene based on its image contents.

Motion compensation is typically used to remove the temporal correlation that exists between frames in an image sequence. Conventional motion compensated prediction methods rely on standard block matching approaches where displacement vectors are estimated over rectangular blocks of the image. This approach is favorable due to its straightforward approach, however, it fails to adequately model object motion which is non-translatory (i.e. object rotation, deformation, or change of scale). This scheme also suffers from annoying blocking artifacts when components of an image feature are assigned different motion vectors. In order to alleviate these problems, several approaches based on digital image warping have been introduced in the past [4, 5, 6, 7]. In these schemes, the predicted frames (also referred to as the current frames) are formed by geometrically transforming or *warping* the previous frames. These methods also fit well within the framework of the object-based coders described earlier.

In this paper, we focus our attention on an object-based approach to motion compensated prediction for videophone-type applications. The techniques of image analysis and digital image warping are utilized to form the predicted frame. This is achieved by segmenting the previous frame into a facial region and a number of arbitrarily shaped regions based on the color information. Each of these regions (represented by a suitable set of control points) are subsequently transformed to form the predicted image. Thus, the motion compensated warping prediction scheme described here consists of three stages: i) Face localization and image segmentation, ii) Control point selection and motion vector assignment, and, iii) Image warping of the previous frame.

2 Face Localization and Image Segmentation

The first step in the coding system described above is the segmentation of the scene into an appropriate set of *regions* or *objects*. Once again, we focus our attention on the specific application of a head-and-shoulders videophone-type sequence. We have found that warping the facial region of the previous frame to obtain the predicted facial region yields unsatisfactory results due to the deformations of the facial features. The opening/closing of the eyes and mouth are failure areas of the warping model due to the covered and uncovered areas created by these deformable features. Thus, we opt to extract the facial area and code this arbitrarily shaped region using a conventional intraframe technique (i.e. Discrete Cosine Transform, Karhunen Loeve Transform, etc.). This will alleviate any coding artifacts within the facial region and provide a more intelligible and perceptually pleasing image to a human observer.

The identification and location of the facial region is determined by utilizing the apriori knowledge of the skin-tone distributions in the perceptual HSV color space. It has been found that skin-colored clusters form within a rather well defined region in the HSV hexcone model for a variety of different skin types [8]. Pixels that fall within a defined polyhedron [8] are classified as skin regions. A set of binary operations which include median filtering, and region filling/removal are used to refine the extracted facial area.

Once the facial area has been identified, then the remaining pixels in the scene must be segmented into a set of distinct regions. A recursive histogramming approach is followed in order to achieve this [9]. Color space conversion is first performed on the RGB image sequences to obtain their corresponding HSV representation. Singularities in the Hue component (i.e. where R,G, and B values are equal) are grouped into different regions according to brightness (i.e. Value) prior to any histogram processing. A histogram of the Hue values is then formed and the most prominent peak is selected. In order to extract these prominent peaks, then the histograms above must be smoothed to remove any meaningless local extrema. For this purpose, we apply the well-known scale space filter [10], where the Hue histogram is convolved with a Gaussian function of zero mean. The peaks and valleys can be determined by examining the first and second derivatives of the convolved result. This process is repeated recursively until no dominant peaks remain in the Hue histogram. After the extraction of each region above, a binary median filter is applied as a post-processing step to smoothen and eliminate small holes within each region, and also to remove any misclassified pixels. A final step in the segmentation process is to classify the remaining pixels into their appropriate regions and merge any regions with a similar hue. A Euclidean distance measure is used in this latter merging step.

3 Control Point Selection & Motion Vector Assignment

Having segmented the image into a set of distinct regions, then a sufficient number of control points must be appropriately selected to represent each region. The partitioning of the image in this way allows the set of regions in the previous frame to be *warped* into their corresponding regions in the current frame. Both, the previous and current frames are available at the encoder while only the former of the two is available at the decoder side. The prediction of the current frame at the decoder is determined by utilizing the previous frame along with the received *warping* instructions (i.e. side information) as follows. The spatial offset (i.e. motion vectors) of the selected control points is determined at the encoder by using the previous and current frames. The encoder finally transmits the previous frame, the computed motion vectors, and the intraframe encoded facial region. The decoder then receives the previous frame along with the necessary side information. The same segmentation and control point selection algorithm as in the encoder are then used to partition the previous frame. As a result, the same control points determined at the encoder are also found at the decoder. These points are spatially shifted to their appropriate positions in the predicted frame according to the received motion vectors. The warping algorithm can finally be applied along with the decoded facial region to obtain the predicted frame. The overhead in the form of side information in this approach consists of the transmitted motion vectors along with the encoded facial area. In [4], the above scheme is known as a forward matching technique and has the advantage that the control points can be selected based on the contents of the image.

A simple way of selecting the control points is by forming a rectangular mesh that partitions the previous frames into a uniform set of non-overlapping blocks. However, a uniform spacing of the selected control points can lead to inaccurate motion vectors which can cause geometric distortions. In order to prevent this from happening, the selection algorithm we use here chooses control points that reside on the edges of the segmented regions determined earlier. In order to simplify the selection algorithm we choose every 10th pixel on each region border and also place a control point at the center of mass of each region. This selection allows each region to be broken up into triangular patches which are formed by the edge points and the center of mass. In this way, each of these patches can be individually *warped* via transformation equations. Control points are also selected at all corners and midpoints of each side of the image frame. These points, however, are stationary and are not spatially offset in the predicted frame. For a better representation of the shape of each region, we also choose control points at pixel boundaries which border three or more distinct regions.

The assignment of motion vectors for each selected control point once again is determined at the encoder where both, the previous and the current frames are available. A modified block matching technique, similar to the one in [4] is also used here, where a 21×21 window size is employed in which the central pixels within this block structure are more heavily weighted. Unlike the scheme in [4], however, the square block used in the matching process here, is not subsampled. Further to this, the mean squared error (MSE) criterion using the Euclidean distance measure is utilized due to the color information. The best match of each selected control point is determined by finding the minimum MSE value within a search space of $+$ or $-$ 15 pixels. These motion vectors are transmitted to the decoder as overhead information along with the encoded facial region.

4 Image Warping

When the decoder receives the previous frame along with the transmitted motion vectors it must predict the current or missing frames. This is accomplished by *warping* the triangles in each of the regions of the previous frame to the corresponding triangles in the predicted frame. The vertices of the triangles in the predicted frame are found by using the same segmentation and control point selection algorithm as in the encoder and spatially shifting these appropriately as mentioned earlier using the motion vector information. Once these are found, then a triangle to triangle mapping follows using an affine transformation [11]. As a result of this, the points within each triangle are geometrically transformed to their corresponding positions. Bilinear interpolation is used when non-integer positions are found. The facial region is finally added to the predicted frame by using the intraframe encoded information. Thus, the current or predicted frame is reconstructed by a number of geometrically transformed patches plus the encoded facial region.

5 Results

The performance of this scheme was evaluated using Frames 100 and 110 of the Claire CIF sequence. A comparison of the warping prediction was made with the conventional block matching motion compensator. In Figures 1a) and 1b) we present Frames 100 and 110 of the original Claire sequence. In Figure 2a), the results of the segmentation and control point selection are shown while Figure 2b) illustrates the performance of the point matching process. As we can see, the extraction of the facial area and segmentation of the remaining regions leads to an appropriate representation of the image contents. Furthermore, we note that the selected control points are tracked quite accurately as shown in 2 b). The standard block matching results are shown in Figure 3a) where 18×18 size blocks were used. The blocking artifacts in this figure are clearly evident in this case where abrupt head rotations are encountered. The transmission overhead required for this 360×288 pixel image amounts to 320 motion vectors. The predicted image using the object-based warping approach is finally shown in Figure 3b) which does not suffer from the annoying blocking artifacts, as in the conventional case. In our object-based approach, a lossless technique has been used to encode the facial region. Only a few minor degradations can be observed in the results of Figure 3 b). The left collar region, and the hair to the right of the facial area are slightly distorted, however, these areas are not visually as important as the critical information conveyed by the facial features. In addition to the improved subjective quality, only 240 motion vectors are required to be transmitted as side information along with the encoded facial area (approximately 500 bits for lossless compression). Future improvements to this technique must focus on a more robust estimation of the motion vectors for the cases of covered and uncovered regions.

6 Conclusions

A object-based warping technique was examined for the temporal prediction of video sequences. In this scheme, a set of control points were determined at the encoder to represent the image contents within a video sequence. These points were selected along the edges of a set of regions that were obtained by a color segmentation method. The facial region was separately encoded to avoid distortions of the facial features. A modified block matching method was then employed at the encoder in order to assign the appropriate motion vectors to the selected points. At the decoder, the same segmentation and control point selection algorithm were used to obtain the control points in the received current frame. These points were subsequently spatially shifted according to the motion vector information. A warping algorithm using an affine transformation was finally used to form the predicted frame. A significant subjective improvement was found in the predicted image using this technique when compared to the conventional block matching approach. Furthermore, the improved visual quality was achieved with only a moderate increase in the transmission overhead and computational complexity of the coding scheme.

References

1. L. Chiariglione, 'MPEG and Multimedia Communications', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 7, No. 1, pp. 5-18, February 1997.
2. H.G. Musmann, M. Hotter, J. Ostermann, 'Object-oriented analysis-synthesis coding of moving objects', *Signal Processing: Image Communication*, Vol. 1, No. 2, pp. 117-138, October 1989.
3. M. Hotter, 'Object-oriented analysis-synthesis coding based on moving two-dimensional objects', *Signal Processing: Image Communication*, Vol. 2, No. 4, pp. 409-428, December 1990.
4. J. Nieweglowski, J. Campbell, P. Haavisto, 'A novel video coding scheme based on temporal prediction using digital image warping', *IEEE Trans. on Consumer Electronics*, Vol. 39, no.3, pp. 141-150, 1993.
5. G. Sullivan, 'Motion Compensation for video compression using control grid interpolation', *IEEE Int. Conf. on ASSP*, pp. 2713-2716, 1991.
6. V. Seferidis, M. Chanbari, 'Generalized block matching motion estimation', *Visual Communications and Image Processing*, SPIE Vol. 1818, pp. 110-119, 1992.
7. J. Nieweglowski, T. Moisala, P. Haavisto, 'Motion compensated video sequence interpolation using digital image warping', *IEEE Int. Conf. on ASSP*, Vol. 5, pp. 205-208, 1994.
8. N. Herodotou, A.N. Venetsanopoulos, 'Image Segmentation for Facial Image Coding of Videophone Sequences', *13th International Conference on Digital Signal Processing*, Santorini, Greece, July 1997.
9. R. Ohlander, K. Price, D.R. Reddy 'Picture Segmentation Using a Recursive Region Splitting Method', *Computer Graphics and Image Processing* Vol. 8, pp. 313-333, 1978.
10. A. Witkin, 'Scale-space Filtering', *Proceedings IJCAI-83*, pp. 1019-1022, Aug 1983.
11. G. Wolberg 'Digital image warping', *IEEE Computer Society Press*, Los Alamitos, California, 1990.

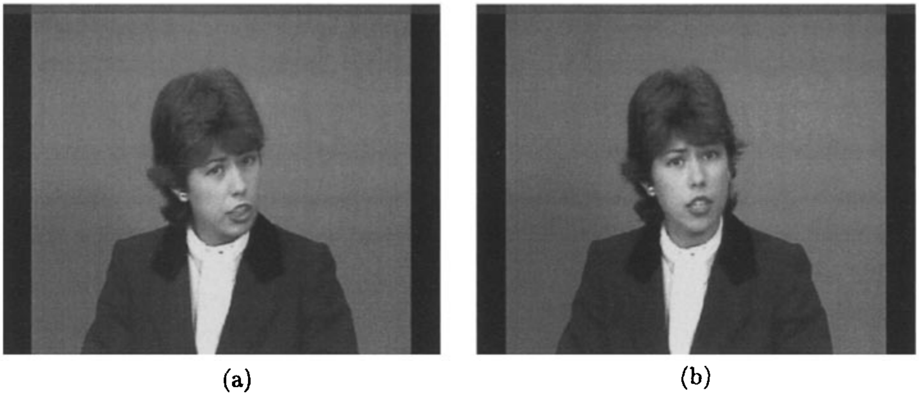


Fig. 1. a) Original Claire, Frame 100, b) Original Claire, Frame 110.

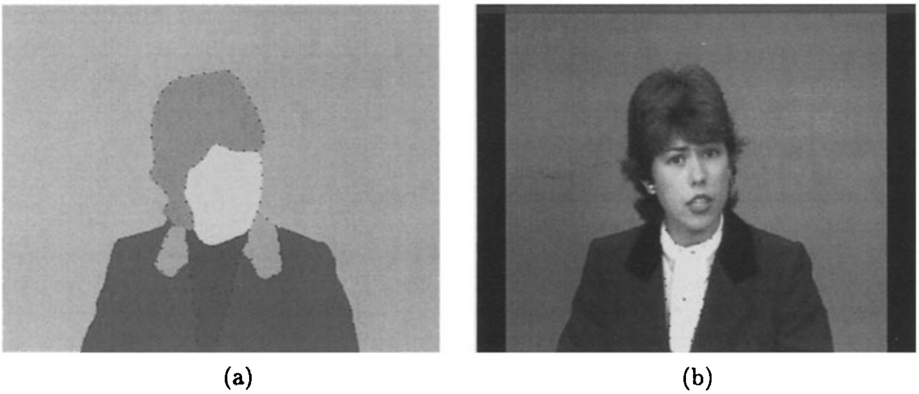
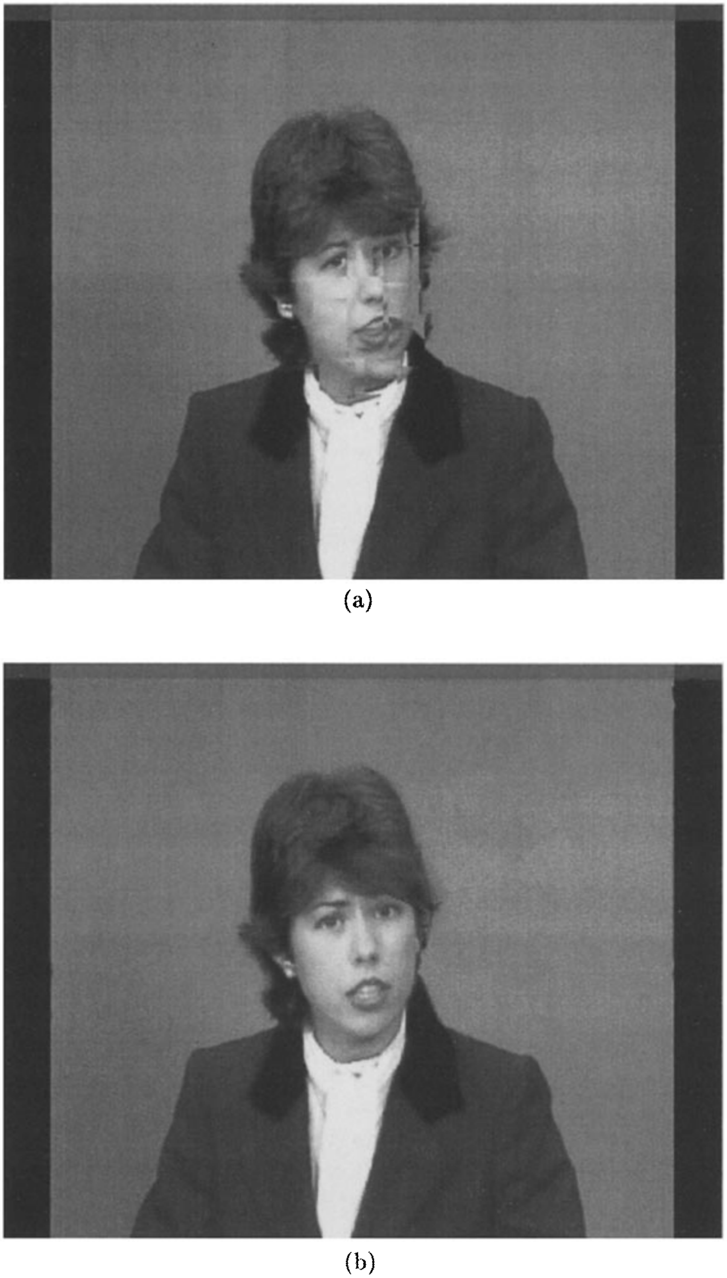


Fig. 2. a) Segmentation and control point selection, b) Matching of control points.



(a)

(b)

Fig. 3. a) Conventional block matching prediction of Frame 110, b) Object-based warping prediction of Frame 110.