An Iterative Technique for 3-D Motion Estimation in Videophone Applications

Gözde Bozdağı[†], A. Murat Tekalp[‡], and Levent Onural[†]
[†]Electrical and Electronics Engineering Department
Bilkent University, 06533 Bilkent, Ankara, TURKEY
phone: +90-312-266-4307 e-mail: bozdagi@ee.bilkent.edu.tr
[‡]Electrical and Electronics Engineering Department
University of Rochester, Rochester, New York, 14627, USA

***Abstract*— In object based coding of facial images, the accuracy of motion and depth parameter estimates strongly affects the coding efficiency. We propose an improved algorithm based on stochastic relaxation for 3-D motion and depth estimation that converges to true motion and depth parameters even in the presence of 50% error in the initial depth estimates. The proposed method is compared with an existing algorithm (MBASIC) in case of different number of point correspondences. The simulation results show that the proposed method provides significantly better results than the MBASIC algorithm.**

1. INTRODUCTION

Image coding is one of the most important problems in image processing since the storage and transmission of digital images requires a very large number of bits. Most work in image coding is based on the fact that any data originated from an image are not random, i.e. adjacent samples exhibit an important spatial correlation. Recently, a new coding method which depends on describing a scene in a higher level sense is beginning to be the prime research topic in image coding [1-5]. This type of coding method is entitled as "object based coding" and represents image signals using structural image models and takes into account the 3-D nature of the scene. The major drawback of this kind of coding is the restriction in the type of the scenes that can be handled. Since dealing with unknown objects is an extremely difficult problem, simplification results if the scene contains a priori known objects. In this way, only the identification of these objects and estimation of their relevant parameters are enough for coding of the scene. Within very low bit rate video communication head and shoulder type scenes are of high interest. Our work is also concentrated on this kind of scenes.

An object based coding system is basically composed of analysis and synthesis parts. A 3-D model of the scene (wire-frame) is utilized at both the transmitter and the receiver sides. 3-D motion and structure estimation techniques are employed at the transmitter to track the motion of the wire-frame model and the changes in its structure from frame to frame. The estimated motion and structure (depth) parameters along with changing texture information are sent and used to synthesize the next frame in the receiver side. So, one of the challenging problems in object based coding of facial image sequences is to adapt a generic wire-frame model developed for an average speaker to fit the actual speaker and to track the 3-D motion of this adapted wire-frame. A general overview of 3-D motion and structure estimation methods can be found in [6]. Among these methods, a point-correspondence method proposed by Aizawa *et al.* [2] have been previously utilized for tracking the motion of the wire-frame once the wire-frame has been fitted manually. This method may not be appropriate for automatic scaling in the z-direction, as it is sensitive to inaccuracies in the initial depth estimates. To this effect, we propose a 3-D motion and structure estimation algorithm utilizing stochastic relaxation. The core of the idea is to add an element of a zero-mean Gaussian or uniform noise to each depth value following the 3-D motion estimation in each iteration. The noise variance is then reduced monotonically as the algorithm progresses. The proposed method is compared with the existing algorithm that is commonly used in object based image coding [2], in case different number of point correspondences.

2. BACKGROUND

In order to estimate the motion in 3-D we have to identify how motion changes the structure of the scene. Let $[X_s(t) \quad Y_s(t) \quad Z_s(t)]^T$ be the vector of the coordinates of a particular point s of a moving object at time t and S refers to the object which is the set of all such points. If we assume that the object is rigid and

0-7803-1772-6/94/$3.00

subject to small rotation, we can express the position of s at time $t+\Delta t$ given its position at time t as,

$$\begin{bmatrix} X_s(t+\Delta t) \\ Y_s(t+\Delta t) \\ Z_s(t+\Delta t) \end{bmatrix} = \begin{bmatrix} 1 & \omega_Z & -\omega_Y \\ -\omega_Z & 1 & \omega_X \\ \omega_Y & -\omega_X & 1 \end{bmatrix} \begin{bmatrix} X_s(t) \\ Y_s(t) \\ Z_s(t) \end{bmatrix} + \begin{bmatrix} T_X \\ T_Y \\ T_Z \end{bmatrix}, \quad \forall s \in S \qquad (1)$$

where ω_X, ω_Y, and ω_Z are the rotational displacements around the X, Y and Z axes, respectively, and T_X, T_Y, and T_Z are the translational displacements along the X, Y and Z axes, respectively. Under orthographic projection along the z-direction, Eq. 1 becomes,

$$\begin{aligned} x_s(t+\Delta t) &= x_s(t) + \omega_Z y_s(t) - \omega_Y Z_s(t) + T_X \\ y_s(t+\Delta t) &= -\omega_Z x_s(t) + y_s(t) + \omega_X Z_s(t) + T_Y, \end{aligned}$$

$$\forall s \in S. \qquad (2)$$

As the only information we can obtain from the 2-D images are the projections of the 3-D objects around us, we have to estimate the rotational and translational displacements from Eq. 2.

In the context of object based coding, we can divide the methods developed for the computation of motion from image sequences into two categories: feature based and optical flow based motion estimation. Among the methods in the literature about feature based motion estimation, MBASIC, recently proposed by Aizawa *et al.* [2], is a simple and effective iterative algorithm for 3-D motion and depth estimation under orthographic projection. MBASIC algorithm requires a set of initial depth estimates which are usually obtained from a generic wire-frame model. Each iteration of the algorithm is composed of two steps: 1) Determination of motion parameters given the depth estimates from the previous iteration, and 2) update of depth estimates using the new motion parameters. Although the performance of MBASIC is very good when the initial depth parameters contain about 10% error or less, it degrades with the increasing amount of error in the initial depth estimates. But in practical applications the initial depth estimates may contain 30% or more error due to problems in scaling the generic wire-frame model to a particular speaker. Thus, in the following section we propose a modification to the MBASIC algorithm which makes it more robust to errors in the initial depth estimates with a small increase in its computational load, thus making it more useful in practical applications. We also compare the performance of the MBASIC algorithm and the improved algorithm in the presence of various degrees of inaccuracy in the initial depth estimates, and show that the improved algorithm converges to the true motion and depth parameters even in the presence of 50% error in the initial depth estimates.

3. PROPOSED METHOD

The proposed method is as follows:

1. Set the iteration counter $m = 0$.
2. Given at least 3 corresponding coordinate pairs $(x_s(t), y_s(t))$ and $(x_s(t+\Delta t), y_s(t+\Delta t))$ and their depth parameters $Z_s(t)$, $s = 1, \ldots, N$, $N \geq 3$, determine the motion parameters using the LSE method given in the MBASIC algorithm.

$$\begin{bmatrix} x_s(t+\Delta t) - x_s(t) \\ y_s(t+\Delta t) - y_s(t) \end{bmatrix} = \begin{bmatrix} 0 & -Z_s(t) & y_s(t) & 1 & 0 \\ Z_s(t) & 0 & -x_s(t) & 0 & 1 \end{bmatrix} \begin{bmatrix} \omega_X \\ \omega_Y \\ \omega_Z \\ T_X \\ T_Y \end{bmatrix} \qquad (3)$$

3. Compute $(x_{s_{(m)}}(t+\Delta t), y_{s_{(m)}}(t+\Delta t))$, the coordinates of the matching points that are predicted by the present estimates of the motion and depth parameters, using Eq. 2. Compute the model prediction error

$$E_m = \frac{1}{N} \sum_{s=0}^{N} e_s \qquad (4)$$

where

$$\begin{aligned} e_s &= (x_s(t+\Delta t) - x_{s_{(m)}}(t+\Delta t))^2 \\ &\quad + (y_s(t+\Delta t) - y_{s_{(m)}}(t+\Delta t))^2. \end{aligned} \qquad (5)$$

Here $(x_s(t+\Delta t), y_s(t+\Delta t))$ are the actual coordinates of the matching points which are given.

4. If $E_m < \epsilon$, stop the iteration,
Else, set $m = m+1$, and perturb the depth parameters as

$$\hat{Z}_{s_{(m)}}(t) \leftarrow \hat{Z}_{s_{(m-1)}}(t) - \beta g(Z_s(t)) + \alpha^m \Delta_s, \qquad (6)$$

where $g(Z_s(t))$ is the gradient of e_s with respect to $Z_s(t)$ (which can be analytically computed from Eq. 5), and, α and β are constants.
For Gaussian distributed perturbations,
$\Delta_s = N_s(0, \sigma^2_{s_{(m)}})$, i.e., zero mean Gaussian with variance $\sigma^2_{s_{(m)}}$, where $\sigma^2_{s_{(m)}} = e_s$.
For uniformly distributed perturbations,
$\Delta_s = U_s(\hat{Z}_{s_{(m-1)}}(t) \pm a_{s_{(m)}})$, i.e., uniformly distributed in an interval of length $2a_s^{(m)}$ about $\hat{Z}_{s_{(m-1)}}(t)$ where U_s denotes uniformly distributed

random numbers. To make reasonable comparisons with the case of Gaussian perturbations, $a_{s_{(m)}}$ is chosen such that

$$\frac{a_{s_{(m)}}^2}{3} = \sigma_{s_{(m)}}^2 = e_s. \quad (7)$$

5. Go to step (2).

3. COMPARISONS

We compare the performance of the MBASIC algorithm and the proposed modified algorithm (with uniform and Gaussian perturbations). The simulations were carried out by using 5, 7 and 10 point correspondences, respectively, with 50% error in the initial depth estimates in each case. The data for the simulations were generated as follows: A set of 5 to 10 points, $(x_s(t), y_s(t))$ with the respective depth parameters $Z_s(t)$, in the range 0 and 1, were arbitrarily chosen. The coordinates $(x_s(t+\Delta t), y_s(t+\Delta t))$ of the matching points in the next frame were generated from $(x_s(t), y_s(t))$ using the transformation (1) with the "true" 3-D motion parameters listed in Table 1. The computed coordinates $(x_s(t+\Delta t), y_s(t+\Delta t))$ are then truncated to the nearest integer. This truncation approximately corresponds to adding 40 dB noise to the matching point coordinates. Then, ±50% error is added to each depth parameter $Z_s(t)$, for the respective simulations. The signs of the error (+ or -) were chosen randomly. At each iteration of the algorithm, first the motion parameters are estimated using the present depth parameters. (This step is the same as in the MBASIC algorithm.) Then, the depth parameters are updated as given by Eq.5. We set $\alpha = 0.95$ and $\beta = 0.3$ to obtain the reported results. In order to minimize the effect of random choices in the evaluation of the results, the results are repeated 3 times using three different seed values for the random number generator. The results shown in Table 1 are the average of these three sets.

Table 1 provides a comparison of the motion parameter estimates obtained by the MBASIC algorithm and the proposed method using uniform and Gaussian distributed random perturbations at the conclusion of the iterations (in this case after 500 iterations). Table 1 shows the results only for the 10-point correspondence case. The 5-point and 7-point results are similar. The comparison of the results of the depth parameter estimation is shown in the figures. In these figures the average estimation error in the depth parameters vs. iteration number is plotted, where the average error is defined as

$$Error = \frac{1}{N}\sqrt{\sum_{i=1}^{N}\frac{(Z_s(t) - \hat{Z}_s(t))^2}{Z_s^2(t)}} \quad (8)$$

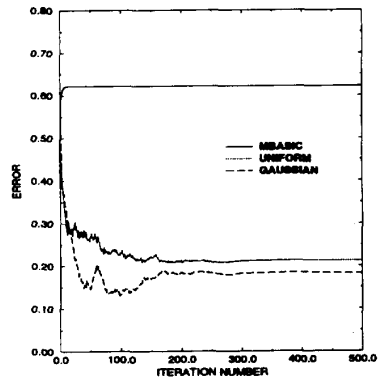
where N is the number of point correspondences; $Z_s(t)$ and $\hat{Z}_s(t)$ are the "true" and estimated depth parameters, respectively. In the MBASIC algorithm, the errors in the depth estimation directly affect the accuracy of the motion estimation and vice versa. This can be seen from Table 1, where the error in the initial depth estimates mainly affects the accuracy of ω_X and ω_Y which are directly multiplied by Z. However, in the proposed algorithm, an update scheme given by Eq. 6 that is indirectly tied to the current estimates of the motion parameters is used. As a result, a smaller average error is obtained for depth parameter estimation. As can be seen from the figures, the depth estimates, using the proposed method, converge closer to the correct parameters even in the case of 50% error in the initial depth estimates. For example, in the case of estimation using 10 point correspondences with 50% error in the initial depth estimates, the proposed method results in about 10% error after 500 iterations whereas the MBASIC algorithm results in 45% error.

	True motion	Aizawa	Uniform	Gaussian
ω_X	0.01	0.0050	0.0107	0.0104
ω_Y	0.02	0.0101	0.0215	0.0209
ω_Z	-0.01	-0.0095	-0.0100	-0.0100
T_X	0.02	0.0154	0.0204	0.0199
T_Y	0.05	0.0523	0.0498	0.0500

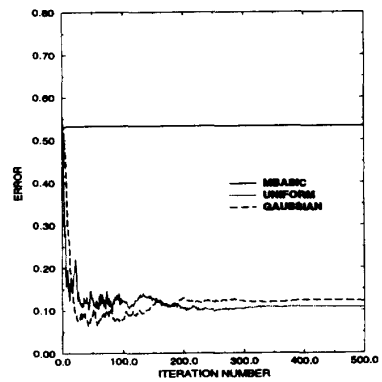
Table 1. The true and estimated motion parameters for 10 point correspondences with 50% initial error in the depth estimates.

3. CONCLUSIONS

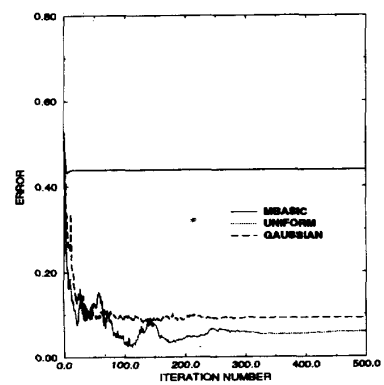
In this paper, we propose an improved motion and structure estimation method that uses point correspondences. We compare our results with those of the basic algorithm proposed by Aizawa *et al.* for different number of point correspondences. It is concluded that the proposed improved algorithm gives better results than MBASIC algorithm and provides a reasonably good performance even in the presence of 50% error in the initial depth estimates. Computational complexity of the improved algorithm is just slightly higher.



(a)



(b)



(c)

Fig 1. Average estimation error in the depth parameters with 50% error in the initial depth estimates for (a) 5, (b) 7, (c) 10 point correspondences.

REFERENCES

[1] R. Forchheimer and T. Kronander, "Image coding-from waveforms to animation," *IEEE Trans. ASSP*, vol. 37, no. 12, Dec. 1989, pp. 2008-2023.

[2] K. Aizawa, H. Harashima, and T. Saito, "Model-based analysis-synthesis image coding (MBASIC) system for a person's face," *Signal Processing: Image Communication*, no. 1, 1989, pp. 139-152.

[3] N. Diehl, "Model-Based Image Sequence Coding," in "Motion Analysis and Image Sequence Processing," M. I. Sezan and R. L. Lagendijk, ed., Kluwer Academic Publishers, 1993.

[4] W. J. Welsh, "Model-based coding of videophone images," *Electronics and Communication Engineering Journal*," Feb. 1991, pp. 29-36.

[5] H. Li, P. Roivainen, and Forcheimer, "3-D Motion Estimation in Model-Based Facial Image Coding," *IEEE Trans. Patt. Anal. Mach. Intel.*, Vol. 15, pp. 545-555, June 1993.

[6] J. K. Aggarwal and N. Nandhakumar, "On the computation of motion from sequences of images - A review," *Proc. IEEE*, vol. 76, no. 8, Aug. 1988, pp. 917-935.