SIMULTANEOUS 3-D MOTION ESTIMATION AND WIRE-FRAME MODEL ADAPTATION INCLUDING PHOTOMETRIC EFFECTS FOR KNOWLEDGE-BASED VIDEO CODING

Gözde Bozdağı[1] *A. Murat Tekalp*[2] *Levent Onural*[1]

[1] Electrical and Electronics Eng. Dept., Bilkent University, 06533 Bilkent, Ankara, TURKEY
[2] Electrical and Electronics Eng. Dept., University of Rochester, Rochester, New York, 14627, USA

ABSTRACT

We address the problem of 3-D motion estimation in the context of knowledge-based coding of facial image sequences. The proposed method handles the global and local motion estimation and the adaptation of a generic wire-frame to a particular speaker simultaneously within an optical flow based framework including the photometric effects of motion. We use a flexible wire-frame model whose local structure is characterized by the normal vectors of the patches which are related to the coordinates of the nodes. Geometrical constraints that describe the propagation of the movement of the nodes are introduced, which are then efficiently utilized to reduce the number of independent structure parameters. A stochastic relaxation algorithm has been used to determine optimum global motion estimates and the parameters describing the structure of the wire-frame model.For the initialization of the motion and structure parameters, a modified feature based algorithm is used. Experimental results with simulated facial image sequences are given.

1. INTRODUCTION

Due to growing interest in very low bit rate digital video (about 10 kbits/s), a significant amount of research has focused on knowledge-based video compression [1]-[5]. Scientists became interested in knowledge-based coding because the quality of digital video obtained by hybrid coding techniques, such as CCITT Rec. H.261 is deemed unsatisfactory at these very low bit rates. Studies in knowledge-based coding employ object models ranging from general purpose 2-D or 3-D models [4, 5] to application specific wire-frame models [1]-[3]. One of the main applications of knowledge-based coding has been the videophone, where scenes are generally restricted to head and shoulder type images. In many proposed videophone applications, the head and shoulders of the speaker is represented by a specific wire-frame model which is present at both the receiver and the transmitter. Then, 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.

Many of the existing methods consider fitting (scaling) a generic wire-frame to the actual speaker using only the initial frame of the sequence [2, 6]. Thus, the scaling in the z-direction (depth) is necessarily approximate. In subsequent frames, global and local motion estimation is usually treated separately, i.e., first the 3-D global motion of the head is estimated under rigid body assumption, and then local motion (due to facial expressions) is estimated making use of action units (AU) [2]. Further, photometric effects (changes in the shading due to 3-D rotations) [7, 8, 9] have not been incorporated into optical flow based motion and structure estimation. Recently, Li *et al.* [3] proposed a method to recover both the local and global motion parameters from the spatio-temporal derivatives of the image. However, they require *a priori* knowledge of the AU's; further they do not consider the photometric effects, nor do they consider the adaptation of the wire-frame model to the speaker.

In this paper, we propose a method where 3-D motion estimation and adaptation of the wire-frame model are considered simultaneously within an optical flow based formulation including the photometric effects. This simultaneous estimation framework is motivated by the fact that wire-frame adaptation, global motion estimation, and local motion estimation are mutually related; thus a combined optimization approach is necessary to obtain the best results. The main contributions of this paper are: (i) photometric effects is included in the optical flow equation, and (ii) the nodes of the wire-frame model are allowed to move flexibly in 3-D, by perturbing the x, y, and z coordinates of the nodes, while constraints about the geometry of the wire-frame model are used to express the effect of the movement of one node on the others. We note here that the adaptation of the wire-frame model from frame to frame serves two purposes which cannot be isolated: (a) to better fit the initial wire-frame model to the speaker in frame $k-1$, and (b) to account for the local motion deformations from frame $k-1$ to frame k without using any *a priori* information about the AU's.

2. INCORPORATION OF PHOTOMETRIC EFFECTS

Since the surface of the wire-frame model is composed of planar patches, the variation in the intensity of a point on the surface due to a change in the normal vector of that patch (e.g., in the case of 3-D rotation) can be computed using a Lambertian surface assumption, and incorporated

0-7803-1775-0/94 $3.00 © 1994 IEEE

in the optical flow equation as [9]

$$I_x u_x + I_y u_y + I_t = \vec{L} \cdot \frac{d\vec{N}(x,y)}{dt}, \tag{1}$$

where I_x, I_y, and I_t are the partial derivatives of the image intensity wrt. x, y and t respectively, u_x and u_y are the 2-D velocities, $\vec{L}$ is the unit vector showing the mean illuminant direction and $\vec{N}$ is the unit normal vector to the surface at point $(x, y, Z(x,y))$. We can express $\vec{N} = (-p, -q, 1)/(p^2 + q^2 + 1)^{1/2}$, where p and q are the partial derivatives of $Z(x,y)$ wrt. x and y respectively.

Under orthographic projection, the 2-D velocities can be expressed in terms of 3-D rotation parameters ω_x, ω_y, ω_z and translation parameters T_x, T_y [10] as

$$\begin{aligned} u_x &= \omega_z y - \omega_y Z + T_x \\ u_y &= -\omega_z x + \omega_x Z + T_y. \end{aligned} \tag{2}$$

Combining (1) and (2), and evaluating $\frac{d\vec{N}(x,y)}{dt}$, we obtain

$$\begin{aligned} I_x(\omega_z y - \omega_y Z + T_x) + I_y(-\omega_z x + \omega_x Z + T_y) + I_t \quad = \\ \vec{L} \cdot \left[\frac{(-p', -q', 1)^T}{\sqrt{p'^2 + q'^2 + 1}} - \frac{(-p, -q, 1)^T}{\sqrt{p^2 + q^2 + 1}} \right] \end{aligned} \tag{3}$$

where $p' = \frac{-\omega_y + p}{1 + \omega_y p}$ and $q' = \frac{\omega_x + q}{1 - \omega_x q}$. The illuminant direction $\vec{L}$ can be estimated from the scene prior to 3-D motion estimation under the assumptions of Lambertian surface, uniform albedo and point light source [11].

3. INCORPORATION OF GEOMETRIC CONSTRAINTS FOR THE WIRE-FRAME MODEL

The depth parameter $Z(x,y)$ for any (x,y) on the i^{th} patch can be expressed as $Z_i(x,y) = (1/c_i)(-p_i x - q_i y + 1)$, where p_i and q_i denote the surface normal of the i^{th} patch (triangle), c_i denote the shift of the patch in the Z direction. Thus, Z can be eliminated from (3).

We can then formulate the problem as to find the parameters $\omega_x, \omega_y, \omega_z, T_x, T_y$ representing the global motion of the wire-frame from frame $k-1$ to frame k, and the parameters $p_i; q_i, c_i$ representing the local adaptation of the wire-frame to minimize the error in (3) over all pixels in frame k. It is important to note that the surface normals of each planar patch on the wire-frame are not independent of each other. Thus, the number of independent unknowns can be reduced by incorporating the structure of the wire-frame in the form of additional constraints

$$p_i x_{ij} + q_i y_{ij} + c_i = p_j x_{ij} + q_j y_{ij} + c_j, \tag{4}$$

where x_{ij} and y_{ij} denote the coordinates of a point that lie on the straight line at the intersection of the i^{th} and j^{th} patches. At each iteration cycle, we visit all the patches of the wireframe model in sequential order. If, at the present iteration cycle, none of the neighboring patches of patch i has yet been visited for updating their structure parameters (e.g., the initial patch), then p_i, q_i, c_i are all independent and are perturbed. If only one of the neighboring patches, say patch j, has been visited (p_j, q_j, c_j have already been updated), then two of the parameters, say p_i and q_i are independent and perturbed. The dependent variable c_i is computed as

$$c_i = p_j x_{ij} + q_j y_{ij} + c_j - p_i x_{ij} - q_i y_{ij}, \tag{5}$$

where x_{ij} is one of the nodes common to both patches i and j, that is either in the boundary or has been already updated in the present iteration cycle. If two of the neighboring patches, say patches j and k, have already been visited, i.e., the variables p_j, q_j, c_j and p_k, q_k, c_k have been updated, than only one variable, say p_i, is independent and perturbed. Then, c_i can be found from eq. 5, and q_i can be evaluated as

$$q_i = \frac{p_k x_{ik} + q_j y_{ik} + c_k - p_i x_{ik} - c_k}{y_{ik}}, \tag{6}$$

where x_{ik} is one of the nodes common to both patches i and k, that is either in the boundary or has been already updated in the present iteration cycle.

The change in the structure parameters p_i, q_i, c_i affects the location of the nodes. At each iteration cycle, the coordinates of a node are updated as soon as the structure parameters of three patches that intersect at that node are updated. Thus, the new X and Y coordinates of the nodes are given by

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} p_i - p_j & p_j - p_k \\ q_i - q_j & q_j - q_k \end{bmatrix}^{-1} \begin{bmatrix} c_j - c_i \\ -c_j + c_k \end{bmatrix}. \tag{7}$$

4. THE PROPOSED METHOD

In the following we summarize the proposed algorithm as:

1. Estimate the illumination direction using the method described in [11].
2. Initialize the coordinates of the nodes (X_n, Y_n, Z_n), for all n, using an approximately scaled initial wire-frame model. Set the iteration counter $m = 0$.
3. Determine the initial motion parameters using the stochastic relaxation method described in [14] (or any other point correspondence method to compute the motion parameters using a set of selected nodes given their depth values).
4. Compute the value of the cost function E given by

$$E = \sum_i \sum_{(x,y) \in i^{th} patch} e_i^2(x,y) \tag{8}$$

where

$$\begin{aligned} e_i(x,y) \quad = \quad & I_x(\omega_z y - \omega_y(p_i x + q_i y + c_i) + T_x) + \\ & I_y(-\omega_z x + \omega_x(p_i x + q_i y + c_i) + T_y) + I_t \\ & -\rho(L_x, L_y, L_z) \cdot \\ & \left(\frac{(-\frac{-\omega_y + p_i}{1 + \omega_y p_i}, -\frac{\omega_x + q_i}{1 - \omega_x q_i}, 1)}{((\frac{-\omega_y + p_i}{1 + \omega_y p_i})^2 + (\frac{\omega_x + q_i}{1 - \omega_x q_i})^2 + 1)^{1/2}} \right. \\ & \left. - \frac{(-p_i, -q_i, 1)}{(p_i^2 + q_i^2 + 1)^{1/2}} \right). \end{aligned} \tag{9}$$

5. If $E < \epsilon$, stop.

Else, set $m = m + 1$, and perturb the motion parameters $\omega = [\omega_x, \omega_y, \omega_z, T_x, T_y]^T$ as

$$\omega_{(m)} \longleftarrow \omega_{(m-1)} + \alpha^m \Delta, \tag{10}$$

where $\Delta = N(0, \sigma^2_{(m)})$, i.e., zero mean Gaussian with variance $\sigma^2_{(m)}$, where $\sigma^2_{(m)} = E$, and

the structure parameters p_i, q_i and c_i as

```
Define count_i as the number of neighbor-
ing patches to patch i whose structure
parameters have been perturbed.
Set count_i=0, for all patches i.

Perturb p_1, q_1, c_1 as
```

$$\begin{aligned} p_{1_{(m)}} &\longleftarrow p_{1_{(m-1)}} + \alpha^m \Delta_1, \\ q_{1_{(m)}} &\longleftarrow q_{1_{(m-1)}} + \alpha^m \Delta_1, \\ c_{1_{(m)}} &\longleftarrow c_{1_{(m-1)}} + \alpha^m \Delta_1, \end{aligned} \tag{11}$$

where $\Delta_i = N_i(0, \sigma_i^{2(m)})$, i.e., zero mean Gaussian with variance $\sigma_i^{2(m)}$, where
$\sigma_i^{2(m)} = \sum_{(x,y)\in patchi} e_i^2(x, y)$.

```
increment count_j, for all j denoting
neighbors of patch 1.

for( i=2 to number of patches)
 {
 if(count_i==1) {
    perturb p_i and q_i
    increment count_m, for all m denoting
    neighbors of patch i.
    Compute c_i.
                }

 if(count_i==2) {
    perturb p_i
    increment count_m, for all m denoting
    neighbors of patch i.
    Compute c_i and q_i.
                }
 If p_i, q_i, and c_i for at least three
 patches intersecting at a node are
 updated, then update the coordinates
 of the node.
 }
```

6. Go to step (4).

Experimental results will be presented in the next section to demonstrate the performance of the proposed method.

	True motion	Initial point	Our method
ω_x	-0.1	-0.08894	-0.1083
ω_y	0.35	0.3368	0.33446
ω_z	-0.03	-0.0113	-0.02683
T_x	6	4.962	5.4719
T_y	-3	-2.8999	-2.7853

Table 1. Global motion estimation without the photometric effects.

5. SIMULATION RESULTS

We have demonstrated the proposed method with synthetic image sequences. The synthetic sequence is generated by moving (the vertices corresponding to the head area) of the textured model of the wire-frame which is an extension of the CANDIDE wire-frame [12] and composed of 217 triangles and 144 nodes [13]. The textured model is obtained by mapping a single frame from the "Miss America" sequence to the initial wire-frame model. The mapping is accomplished after scaling the wire-frame model approximately to the location and the size of the face by positioning four extreme points, interactively. A second frame is obtained from the first one by rotating and translating it. We also include the local motion specified by the AU2, AU17 and AU46 which correspond t o outer brow raiser. Then we applied our algorithm to check its performance in finding these already known motion parameters. The results of global motion estimation with no photometric effects is presented in Table 1. In addition, we have synthesized a new frame from the first frame using the estimated motion parameters and computed the difference between this synthesized frame and the second frame. The mean square synthesis error between the actual second frame $I_a(x, y)$ and the synthesized second frame $I_s(x, y)$ is computed according to

$$MSE = \left(\frac{1}{N \times M} \sum_{x,y} (I_a(x, y) - I_s(x, y))^2 \right)^{\frac{1}{2}}, \tag{12}$$

where N and M show the x and y extends of the image, respectively. The MSE in this case is 7.554. We repeat the above experiment by including the photometric effects. Table 2 shows the true and estimated global motion parameters in this case. The MSE between the second and the synthesized frames is now equal to 7.101. The first, second, and the synthesized frames for the both cases are shown in Figs. 1 and 2, respectively.

6. CONCLUSION

In this paper, we address the problem of 3-D motion estimation in the context of knowledge-based coding of facial image sequences. Our experiments show that the simultaneous estimation gives more accurate results than the ones found in the literature [2],[3]-[4] as expected due to the mutuality of the estimated parameters. The incorporation of the photometric effects to the formulation also improves the estimation results by a considerable amount. Future work at this point will include analysis of the quantization effects

	True motion	Initial point	Our method
ω_x	-0.1	-0.08894	-0.1079
ω_y	0.35	0.3368	0.34001
ω_z	-0.03	-0.0113	-0.0272
T_x	6	4.962	6.4660
T_y	-3	-2.8999	-2.8852

Table 2. Global motion estimation including the photometric effects.

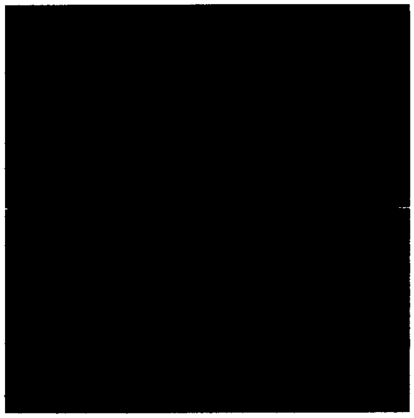

Figure 1. (a) The first frame of "Miss America", (b) simulated second frame with global and local motion (without photometric effects); (c) synthesized second frame using the estimated motion and structure parameters.

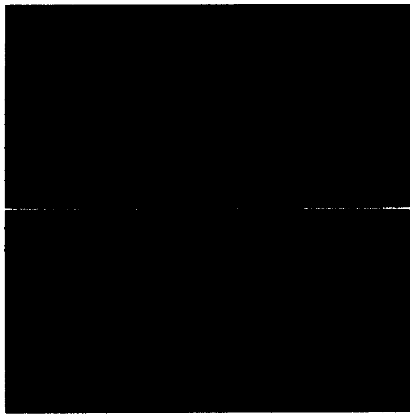

Figure 2. (a) The first frame of "Miss America", (b) simulated second frame with global and local motion,and the photometric effects; (c) synthesized second frame using the estimated motion and structure parameters.

to the performance of the algorithm and implementation of the parameter coding. Also, to decrease the synthesis error further, the proposed method is aimed to be modified to take care of the change in texture as a result of motion.

REFERENCES

[1] R. Forchheimer and T. Kronander, "Image Coding - From Waveforms to Animation," *IEEE Trans. Acoust. Speech Sign. Proc.*, vol. ASSP-37, no. 12, Dec. 1989, pp. 2008-2023.

[2] K. Aizawa, C. S. Choi, H. Harashima, and T. S. Huang, "Human Facial Motion Analysis and Synthesis with Application to Model-Based Coding," *in* Motion Analysis and Image Sequence Processing, (M. I. Sezan and R. L. Lagendijk, eds.), Kluwer Academic Publishers, 1993.

[3] H. Li, P. Roivainen, and R. Forcheimer, "3-D Motion Estimation in Model-Based Facial Image Coding," *IEEE Trans. Patt. Anal. Machine Intel.*, vol. PAMI-15, no. 6, Jun. 1993, pp. 545-555.

[4] J. Ostermann, "An Analysis-Synthesis Coder Based on Moving Flexible 3D-Objects," *Proc. Pic. Coding Sym.*, Lausanne, Switzerland, Mar. 1993.

[5] 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.

[6] M. J. T. Reinders, B. Sankur, and J. C. A. van der Lubbe, "Transformation of a General 3D Facial Model to an Actual Scene Face," *11th Int. Conf. Pattern Recog.*, 1992, pp.75-79.

[7] A. Pentland, "Photometric Motion," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. PAMI-13, no. 9, Sep. 1991, pp. 879-890.

[8] A. Verri and T. Poggio, "Motion Field and Optical Flow: Qualitative Properties," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. PAMI-11, no. 5, May 1989, pp. 490-498.

[9] J. N. Driessen, Motion Estimation for Digital Video, Ph.D. Thesis, Delft University of Technology, Dept. of Electrical Eng., Delft, The Netherlands, Sept. 1992.

[10] B. Klaus and P. Horn, Robot Vision, MIT Press, 1986.

[11] Q. Zheng and R. Chellappa, "Estimation of Illuminant Direction, Albedo, and Shape from Shading," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. PAMI-13, no. 7, July 1991, pp. 680-702.

[12] M. Rydfalk, "CANDIDE: A parametrised face," Dept. Elec. Eng. Rep. LiTH-ISY-I-0866, Linköping Univ., Oct. 1987.

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

[14] G. Bozdağı, A. M. Tekalp, and L. Onural, "Improved 3-D Motion and Depth Estimation using Stochastic Relaxation for Video phone Applications," submitted to the IEEE Trans. Image Proc., Special Issue on Image Sequence Processing, April 1993.

*This research has been partially supported by NATO International Collaborative Research Grant N 900012 and by TÜBİTAK under the COST 211 project.