8—32

Global motion estimation based on the constrained spatio-temporal gradient method in model-based coding

Young Wook Sohn[1], Doo-Hyun Kim[1], Dong-O Kim[1], and Rae-Hong Park[1]
[1]Department of Electronic Engineering, Sogang University

ABSTRACT

For global motion estimation in model-based coding, this paper proposes a constrained spatio-temporal gradient method using contour information. To overcome the local minimum problem in motion tracking of the conventional spatio-temporal gradient method, the translational matching position detected based on 2-D contour matching and the rotational motion model are introduced as constraints. Simulation results show that the proposed method yields smaller mean square error (MSE) than the conventional method.

1. INTRODUCTION

Model-based coding is one of approaches that parameterize an object, e.g., head or facial component in a head-and-shoulder image. Assuming a parameter-driven wireframe model of a face, an encoder analyzes motion parameters of a face and transmits them, and a decoder synthesizes a facial image based on them. If an encoder can parameterize an object accurately, the amount of bits for representation of parameters is very low. In this term, model-based coding has been studied as a very low bit-rate synthetic image coder and its accurate motion parameterization has been primarily investigated. For motion estimation and image synthesis, fitting of a wireframe to the first frame (initialization) is also needed, and the texture of the first frame arid the changed parts in subsequent frames should be transmitted. There has been few general method to initialize a wireframe, which is one of the main concern in model-based coding. A texture transmission method [1-3] uses the traditional image coding method, i.e., discrete cosine transform (DCT). Recently, model-based coding is regarded as an application to man-machine interfacing and virtual reality, and one of the most important parts of model-based coding is the accurate and robust estimation of parameters.

As a human face contains various facial expressions, it is difficult to efficiently parameterize a face with a simple 2-D model. So 3-D analysis and synthesis of a facial image are needed and there are famous 3-D wirefame models like Candide [4], which is controlled by wireframe parameters. The wireframe motion parameters are divided into two parts: global and local. Global motion parameters with translational arid rotational values specify the 3-D position and orientation of the wireframe model, while local ones describe the facial expressions of the model. Facial expressions are represented by action units (AUs) [5] that describe various movement combinations of facial components such as eyes, eyebrows, and mouth. If global motion parameters are found inaccurately, the local ones cannot be estimated correctly because some facial components are mislocated. In real image sequences, finding global motion parameters is difficult because they depend on local motions. Also in the context of global motions, local motions and noise cannot be effectively distinguished, which is one of difficulties in motion estimation.

To find global motion parameters quickly and accurately in model-based coding, the spatio-temporal gradient methods have been presented [3,4]. Also in computer vision field, feature tracking and contour matching methods were developed [6,7]. The spatio-temporal gradient method computes 3-D parameters by combining the orthogonal/perspective projection model with the optical flow constraint, in which least-square formulation is used. The spatio-temporal gradient method is quite accurate for small motions and noise. Utilizing feature points or contour information can lead to robust algorithms. However, it is quite difficult to find the corresponding feature points in the previous image if occlusion occurs. Contour information can be found around mouth, eyes, and chin. Kass *et al.* [8] proposed 'snakes', with the form of active contours. Contours can robustly find the locations of the facial components. However, it is not appropriate to directly apply this 2-D approach to 3-D environments. In this paper, a constrained spatio-temporal gradient method for global motion estimation is proposed. The reference point obtained by 2-D contour matching is used in the context of the least-square formulation, and the error computed at each step is used for weighting the constraint.

2. FACIAL REGION TRACKING BASED ON THE SPATIO-TEMPORAL GRADIENT METHOD

The block diagram of a typical model-based coder is shown in Fig. 1. It is composed of three parts in both the encoder and decoder: global motion estimation, local motion estimation, and texture update. Additionally, the initialization part is used for the first frame of the sequence. At the encoder and decoder, a wireframe model is shown and motion analysis is based on the model. In this paper, only the wireframe model, global motion

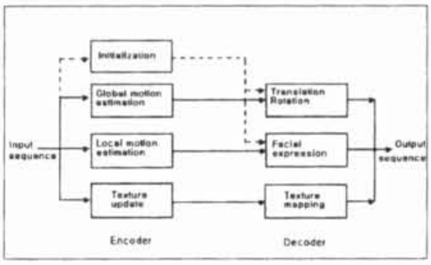


Fig. 1. Block diagram of a typical model-based coder.

[1] C.P.O. Box 1142, Seoul, Korea 100-611. E-Mail: chopin42@samsung.ac.kr, dhkim@eevision1.sogang.ac.kr, hyssop@eevision1.sogang.ac.kr, rhpark@ccs.sogang.ac.kr

analysis, and texture mapping are described. The main purpose of this paper is to estimate robust global motion parameters by neglecting any local motion parameters and texture update.

A. Wireframe model

In model-based coding, it is assumed that a wireframe model is known at a transmitter and receiver. An example of a wireframe model is shown in Fig. 2(a), where the model is an approximate shape of a face. The position of the model is set by global parameters and its facial expression is controlled by local parameters that describe the movement of each facial component. Fig. 2(b) illustrates a wireframe showing local expression of opening the mouth. The model is constructed according to the shape in [3]. About 300 triangle patches are used in this model. If more triangles are used, the shape will be more accurate and natural, which increases the computational cost. Because a matching model is known, approximated depth information can be obtained from the wireframe model specified.

B. Conventional spatio-temporal gradient method for global motion estimation

A conventional spatio-temporal gradient method is based on the following rigid-body assumption;

$$I(x,y,t) = I(x+a, y+b, t+1) \tag{1}$$

where $I(x,y,t)$ and $I(x,y,t+1)$ represent intensities of a pixel at (x,y) at time t and $t+1$, respectively, and a and b denote the translational motion along x and y axes, respectively.

The assumption signifies that the intensity of a pixel does not vary in time. Using Taylor series expansion and some rearrangements give the first-order optical flow constraint [9]

$$I_x u + I_y v + I_t = 0 \tag{2}$$

where $I_x=\partial I(x,y,t)/\partial x$, $I_y=\partial I(x,y,t)/\partial y$, and $I_t=\partial I(x,y,t)/\partial t$ represent derivatives of $I(x,y,t)$ with respect to x, ,y, and t, respectively. Velocities along x and y directions are denoted by $u=\partial x/\partial t$ and $v=\partial y/\partial t$, respectively, and (u,v) is referred to as a 2-D motion vector of a point. Derivatives I_x, I_y and I_t in a digital image are approximated by $I_x=I(x+1,y,t)-I(x,y,t)$, $I_y= I(x+1,y,t)-I(x,y,t)$, and $I_t= I(x+1,y,t)-I(x,y,t)$, respectively.

To obtain 3-D parameters, the 2-D motion vector (u,v) is described in terms of 3-D motion parameters. If an orthogonal projection model is used, x and y components of (u,v) at (x,y), denoted by $u_{x,y}$ and $v_{x,y}$, can be represented as

$$\begin{aligned} u_{x,y} &= w_z y - w_y z_{x,y} + T_x \\ v_{x,y} &= -w_z x + w_x z_{x,y} + T_y \end{aligned} \tag{3}$$

where w_x, w_y and w_z are rotational parameters about the x, y, and z axes, respectively, and T_x and T_y are translational parameters along the x and y directions, respectively. Another unknown value $Z_{x,y}$, the depth information at (x,y) is obtained from the given wireframe model.

Inserting Eq. (3) into Eq. (2) yields [3]

$$\begin{bmatrix} -z_{x1,y1}I_{y1} & z_{x1,y1}I_{x1} & x_1 I_{y1} - y_1 I_{x1} & I_{x1} & I_{y1} \\ -z_{x2,y2}I_{y2} & z_{x2,y2}I_{x2} & x_2 I_{y2} - y_2 I_{x2} & I_{x2} & I_{y2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -z_{xN,yN}I_{yN} & z_{xN,yN}I_{xN} & x_N I_{yN} - y_N I_{xN} & I_{xN} & I_{yN} \end{bmatrix} \begin{bmatrix} w_x \\ w_y \\ w_z \\ T_x \\ T_y \end{bmatrix} = \begin{bmatrix} -I_{t1} \\ -I_{t2} \\ \vdots \\ -I_{tN} \end{bmatrix} \tag{4}$$

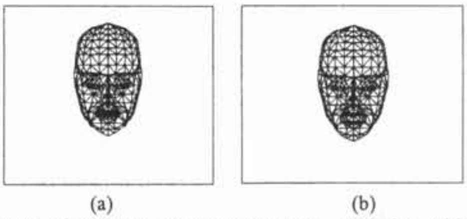


Fig. 2. Wireframe model: (a) facial wireframe, (b) facial wireframe with mouth open.

where N denotes the number of pixels considered, and I_{xn}, I_{yn} and I_{tn} represent I_x, I_y, and I_t at the n^{th} point, respectively. Eq. (4) can be expressed, equivalently, in vector-matrix form

$$\mathbf{A}X = Y \tag{5}$$

where $\mathbf{A}$ denotes an $N\times5$ matrix, X is a 5×1 parameter vector, and Y represents an $N\times1$ temporal-gradient vector. Then the parameter vector $\mathbf{X}$ is expressed as

$$X = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T Y \tag{6}$$

where the superscript -1 denotes inverse operations.

Since this least-square method uses a number of samples in a facial area that contains various changes of expressions, a number of outliers may exist among the extracted samples. Sampling only in unchanged area of a face reduces the number of outliers, which is not easy practically. Or, outliers such as the pixels related to facial expression components or to noise must be rejected. Many methods have been presented. Li and Forchheimer used M-estimators to reject outliers [10], and Choi *et al.* [3] rejected outliers error values of which are greater than the average error. However these methods cannot guarantee that their solutions are global minima or at least they are good for the human visual system (HVS), which is the limitation of the conventional spatio-temporal gradient method. This approach is good at finding accurate motion parameters in the limited range (up to about 10 pixels), but after long sequence tracking, sometimes its results degrade greatly because its solution falls into local minima, resulting in deviated face boundary tracking. This case yields poor subjective evaluation though it produces relatively small MSE, compared with other tracking techniques. The MSE defined by

$$MSE = \frac{1}{K}\frac{1}{L}\sum_{k=1}^{K}\sum_{l=1}^{L}(I(k,l) - I'(k,l))^2$$

is minimized by the least-square method, where $I(k,l)$ ($I'(k,l)$) represents the original (synthesized) pixel intensity and the synthesized pixel intensity in $K\times L$ area of interest that is smaller than the entire image size. Since Eqs. (1)-(4) correspond to modeling based on minimization of the MSE, this approach gives small MSE. But in long sequences, though the MSE is continuously minimized, filling in local minima may occur because of abrupt motions or large noise. Note that minimization of the MSE does not always lead to best solutions related to the HVS.

C. Synthesis of a facial image

After the model parameters are found, they are transmitted and the receiver synthesizes a facial image using them. Assuming that the texture of the first frame is already known at a receiver, each triangle patch in the model is plated with the texture information by means of the affine transform:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{11} \\ a_{21} & a_{22} & a_{22} \end{bmatrix} \begin{bmatrix} x & y & 1 \end{bmatrix}^T$$

where a_{ij}, $1 \le i \le 2$, $1 \le j \le 3$, denotes the affine transform coefficients, the superscript t means transposition, and (x,y) ((x',y')) signifies the original (transformed) position.

3. PROPOSED GLOBAL MOTION ESTIMATION

To overcome the local minimum problem, several numerical methods have been presented. One of them is simulated annealing that tries all possible positions gradually. It is a useful numerical method, however it requires a high computational complexity. In head tracking, some cues can be used to overcome the local minimum problem, without trying all possibilities. As mentioned before, some computer vision methods, i.e., feature tracking or contour matching, can approximately find the position of a head. Although it gives a larger MSE than the spatio-temporal gradient method, it produces reasonable results in terms of subjective performance related to the HVS. So the useful information from other approaches can be combined with the conventional spatio-temporal gradient method to avoid local minima. Some values obtained from other approaches are used as constraints in solving the least-squares formulation with a Lagrangian multiplier. Contour matching, based on the concept of 'snake' [8] is employed in the proposed algorithm. The 2-D position that maximizes the contour energy $E_{contour}$, defined by

$$E_{contour} = \int_c (E_{line} + E_{edge}) ds \tag{7}$$

is used as the matching point of the contour, where the line energy $E_{line}=I(x,y)$ and edge energy $E_{edge}=-|\triangledown I(x,y)|^2$ are defined by the pixel intensity and the negative of the gradient magnitude square of $I(x,y)$, respectively, with the symbol $\triangledown$ representing the gradient operation. In this paper, the contour consisting of chin boundaries of the wireframe model is used as the reference contour and the position that maximizes Eq. (7) is detected.

The 2-D (x,y) parameter of the wireframe model is detected by searching the matching point. This matching point detected can be denoted as the translational vector $X_{contour}=[x_{contour}\ y_{contour}]^t$ of the reference contour.

Besides the translational vector $X_{contour}$, obtained from contour matching, another constraint for rotational parameters is assumed. In long sequence tracking, rotational parameters may change about three axes. But after a while, in most cases, rotational parameters return to the values that represent the frontal directions with all the rotational parameters equal to zeros. So this assumption is modeled and also used as an additional constraint. This modeling can be constructed in terms of the rotational vector $X_{rotation}$ expressed as $X_{rotation}=[0\ 0\ 0]^t$. If the constraint parameter vector $X_c = [X_{rotation}, X_{contour}]^T$ is used, the constrained least-squares equation fur the constrained cost function J_c with a Lagrangian multiplier α can be expressed as

$$J_c(X) = ||\mathbf{A}X-Y||^2+\alpha||\mathbf{T}X-X_c||^2 \tag{8}$$

where $||\cdot||$ represents a norm of a vector. The nondiagonal components in the matrix $\mathbf{T}$ denote the dependency between a pair of parameters, which are set to zeros in our experiments. The constraint parameter vector X_c consists of the translational parameter vector $X_{contour}$, obtained from 2-D contour matching and rotational parameter vector $X_{rotation}$ by the rotational model. As α increases, the solution of Eq. (8) becomes closer to that solely determined by the constraint. In experiments, α is set to 100. The matrix $\mathbf{T}$ and the constraint parameter vector X_c are expressed as

$$\mathbf{T}=\begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} & t_{15} \\ t_{21} & t_{22} & t_{23} & t_{24} & t_{25} \\ t_{31} & t_{32} & t_{33} & t_{34} & t_{35} \\ t_{41} & t_{42} & t_{43} & t_{44} & t_{45} \\ t_{51} & t_{52} & t_{53} & t_{54} & t_{55} \end{bmatrix},$$

$$X_c = \left[X_{rotation} \quad X_{contour}\right]^t = \left[0 \quad 0 \quad 0 \quad x_{rotation} \quad y_{contour}\right]^t$$

where $\mathbf{T}$ is set to the 5×5 identity matrix $\mathbf{I}_5$ in experiments.

Differentiation of Eq. (8) with respect to X and setting to zero gives

$$X=(\mathbf{A}^T\mathbf{A}+\alpha\mathbf{T}^T\mathbf{T})^{-1}(\mathbf{A}^TY+\alpha\mathbf{T}^TX_c). \tag{9}$$

Note that the minimum MSE criterion sometimes yields the solution that is not desirable in terms of subjective evaluation. If the face part of the wireframe model tracked by the conventional least-squares method is significantly misaligned, it is deairable to discard that solution and try to escape the local minimum point. In the proposed algorithm, iterative weighting is employed to strongly constrain against the local minima. A weighting coefficient e_k is determined in terms of the average squared error computed at the k^{th} iteration:

$$e_k = \beta\left\|\mathbf{A}X^k - Y\right\|^2 \Big/ N \tag{10}$$

where X^k denotes the parameter vector computed by Eq. (9) at the k^{th} iteration and β is a constant set to 200 in experiments.

Applying this weighting coefficient to the constraint term gives the weighted cost function J_w defined by

$$J_w(X)=\|\mathbf{A}X-Y\|^2-\alpha e_k\|\mathbf{T}X-X_c\|^2 \tag{11}$$

and it gives the solution:

$$X=(\mathbf{A}^T\mathbf{A}+\alpha e_k\mathbf{T}^T\mathbf{T})^{-1}(\mathbf{A}^\mathbf{T}Y+\alpha e_k\mathbf{T}^TX_c). \tag{12}$$

4. EXPERIMENTAL RESULTS

The 288×352 Miss America image sequence consisting of 150 frames is used in experiments. The first frame of the test sequence is shown in Fig. 3. The performance of the proposed method is compared with that of the conventional spatio-temporal gradient method [3] in terms of the MSE. In tracking the sequence, five iterations are performed for both methods. Note that only global motions are estimated and local motion (facial expressions) analysis is not considered here. Local motions are considered as noise and the corresponding pixels are regarded as outliers and rejected. Outliers are rejected in both methods if their error values are greater than the average error [3]. The initial position of the wireframe is fitted manually and an example of initial setting is shown in Fig. 4.

Fig. 5 shows the MSE as a function of the frame number by the conventional spatio-gradient method and the proposed method. In some frames, the MSE of the proposed method is smaller than that of the conventional method. The MSE values by both methods are almost the same in the frames of which MSEs are relatively small. In Eq. (12), large weighting is employed as constraints for the case in which the error in Eq. (10) from the previous iteration is relatively large. In the frames that produce a relatively large MSE, e.g., larger than 20, the proposed method reduces the MSE value by imposing the constraint greatly. But in some frames, the MSE by the proposed method is larger than that of the conventional method, where the contour information does not lead to accurate matching.

A tradeoff exists between the MSE and subjective evaluation: MSE values can be quantitatively computed while there is no general measure for subjective evaluation. So new measures faithfully reflecting subject performance are needed. Locus of (x,y) positions tracked by the motion estimation algorithm can be one of them, because it is easily perceived to human eyes. In experiments, the position detected by 2-D contour matching is assumed to be appropriate to the HVS.

Fig. 6 shows loci of (x,y) positions detected by the conventional and proposed algorithms relative to the positions detected by contour matching. Contour matching is considered to be a representation appropriate for the subjective evaluation. The closer the locus to the center of the graph, the smaller error it gives, compared with the locus detected by contour matching. The start positions of both methods are at the origin, and the final positions are close to each other. For simple representation, the two loci are plotted every five frames. The overall scope of the locus by the proposed method is closer to the center of the graph than that of the conventional method, which means that the tracking trajectory by the proposed method closely follows that of the contour information.

Fig. 7 shows the wireframe superimposed on the input image (82nd frame). Figs. 7(a) and 7(b) show results of the conventional and proposed methods, respectively. Fig. 7(a) shows the misaligned wireframe caused by the local minimum problem whereas Fig. 7(b) shows the reasonable wireframe tracking. Performance comparison with Fig. 7 shows that for the cases in which the spatio-temporal method fails to track the input image sequence, the contour information can be used as a useful cue. When this information is formulated as a weighted constraint, it has a noticeable effect on the frames whose MSE values are relatively large.

5. CONCLUSIONS AND FUTURE WORKS

To overcome the local minimum problem in motion tracking of the conventional spatio-temporal gradient method, the proposed algorithm introduces additional informations such as the translational parameters from contour matching and the rotational parameters from the motion modeling. Simulation results show that the proposed global motion estimation algorithm based on the constrained spatio-temporal gradient method yields noticeable image quality improvement for significantly misaligned cases of motion tracking. Further research will focus on development of more accurate and reliable cues helpful for detection of the global minimum.

Fig. 3. Test frame.

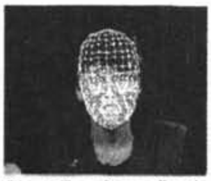


Fig. 4 Test frame fitted by initial position.

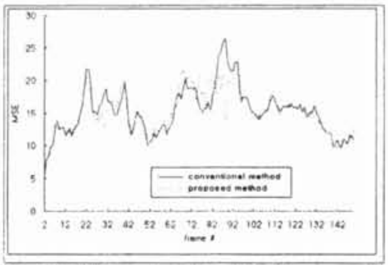


Fig. 5. MSE as a function of the frame model.

REFERENCES

[1] Y. Nakya, K. Aizawa, and H. Harashima, "Texture updating methods in model-based coding of facial images," in *Proc. Picture Coding Symposium* '90, Cambridge, MA, 731-732 (1990).

[2] T. S. Huang and L. Tang, "Model-based video coding - Some challenging issues," in *Multimedia Communications and Video Coding*, Y. Wang *et al.*, Eds., Plenum Press, 215-222 (1995).

[3] C. S. Choi, K. Aizawa, H. Harashima, and T. Takebe, "Analysis and synthesis of facial image sequences in model-based image coding," *IEEE Trans. Circuits and Svstem for Video Tech.*, vol. CSVT-4, 257-275 (Jun.1994).

[4] H. Li, P. Roivainen, and R. Forchheimer, "3D motion estimation in model-based facial image coding," *IEEE Trans. Pattern Analysis and Machine Intelligence* PAMI-15, 545-555 (Jun.1993).

[5] P. Ekman and W. V, Friesen, *Facial Action Coding System*, Consulting Psychologist Press Inc. (1977).

[6] A. Lanitis, C. J. Taylor, and T. F. Cootes, "A unified approach to coding and interpreting face images," in *Proc. Int. Conf. Computer Vision '95*, Cambridge, MA, 368-380 (1995).

[7] M. J. Black and Y. Yacoob, "Tracking and recognizing rigid and non-rigid facial motions using local parametric model of image motion," in *Proc. Int. Conf. Computer Vision '95*, Cambridge, MA, 374-381 (1995).

[8] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," in *Proc. Int. Conf. Computer Vision '87*, London, England, 259-268 (1987).

[9] B. K. Horn and B. G. Schunck, *Robot Vision*, MIT Press (1986).

[10] H. Li and R. Forchheimer, "Two-view facial movement estimation," *IEEE Trans. Circuits and Svstems for Video Technology*, vol. CSVT-4, 276-287 (Jun.1994).

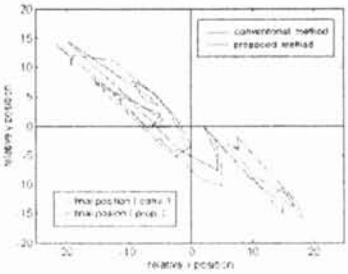
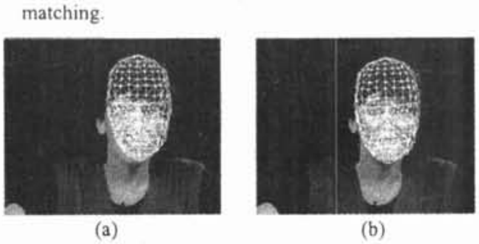


Fig. 6. Loci of positions detected by conventional and proposed algorithms relative to postions detected by contour matching.



(a) (b)

Fig. 7. Wireframe superimposed on the input image: (a) conventional method, (b) proposed method.