SEVERAL KEY PROBLEMS IN MODEL-BASED IMAGE SEQUENCE COMPRESSION BY USING INTERFRAME AUs CORRELATION

Defu Cai Xiangwen Wang Huiying Liang

Institute of Electronics, Chinese Academy of Science
17 Zhong Guan Cun Road, Beijing 100080, China
Tel: 0086-1-2554459 FAX: 0086-1-2567363

ABSTRACT

The model-based image coding has received much attentions for it could gain a big compression ratio and well form the basis of very low bit rates visual transmissions. Many authors had contributed very much to this area. However, some key problems are not well solved for practice. In this paper, the human face model and its invariant representation, textural mapping for natural synthesizing of facial images, occlusion problem in case of face rotation, real time automatic extraction of face feature points, a better performance criteria suitable for facial expression analysis / synthesis are discussed. The results carried by computer simulation are given.

1. INTRODUCTION

The model-based image coding (MBIC) has received much attentions for it could gained a big compression ratio and well form the basis of very low bit rates visual transmissions. Only a limited analyzing information have to be sent because of restricted scene of head-and-shoulder and an explicated head model used. During the last decade, a great deal of progress has been made in MBIC. Many authors had contributed very much to this area[1-11]. However, some key problems are not reasonable solved yet for practice. As known, efficient object (human face in general) representation is a key to the successful face modeling in MBIC. Representations, which are invariant to imaging process and must be not sensitive to changes in the appearance of a human face, are represented as a group of transformations. Invariant representations do not require the computation of camera altitude or pose and therefore promises significant computation advantages. For instance, the rigid invariant representation gain the benefits for eliminating the need of computing the orientation of the face respected to camera. In other hand, invariant representation make the possibility to integrate informations across different image frame. This integration offers robustness and stability in building a model-based image sequence encoder, especially in making use of the interframe AUs correlation and other properties from multiple or successive image frames.

Textural mapping is a powerful tool for facial image synthesis. In the case of wireframe model, three vertices of a triangle patch in texture plane (image) could mapped to the new ones of deformable patch (screen). How to form a natural synthesized facial image rely on the textural mapping of every points located in the patches.

The judgment of how consistency of two image versions before and after facial expression analysis/synthesis is the problem remained. The performance criteria used in existing image coding, such as MSE, SNR etc. are not suitable well for MBIC. A new performance criteria should be presented urgently in near future.

In this paper, several key problems in MBIC are discussed. They are including the human face model and its invariant representation, textural mapping for natural synthesizing of facial images, occlusion problem in case of face rotation, real time automatic extraction of face feature points and a better performance criteria suitable for facial expression analysis / synthesis are discussed. The results carried by computer simulation are given.

2. HUMAN FACE MODEL AND ITS 3-D INVARIANT REPRESENTATION

2.1. Constrained nonrigid motion model

For the human face is a deformable elastic object, the motions of live face are highly nonrigid. Parameterizing for animating and synthesizing facial expressions is a rigorous task. We build on our dynamic motion model the complex motions that incorporate the rigid (global) and nonrigid (local deformable tissues) into geometric modeling primitives. In visual application cases, the face motions are smoothly frame by frame. This is an important constraint, which diminish the ill pose problem. Then, we built the motion model for generalized face motion according to Helmholz theory:

$$p' = G(p) + \Delta(p)$$

where $p = (x,y,z)^T$ and $p' = (x',y',z')^T$ denote the position vector of the points in the face before and after motion.

(1). global (rigid) motion

When four points $\mathbf{p}_i = [X_i,Y_i,Z_i]^T$, $\mathbf{p}_j = [X_j,Y_j,Z_j]^T$, $\mathbf{p}_k = [X_k,Y_k,Z_k]^T$ and $\mathbf{p}_0$ the original are adopted in a special frame, we have:

0-8186-6950-0/94 $4.00 © 1994 IEEE

$$\begin{aligned} &x^T G^{-1} y = 0 \\ &x^T G^{-1} x - y^T G^{-1} y = 0 \end{aligned} \qquad (1)$$

and

$$G = \begin{bmatrix} p_i^T p_i & p_i^T p_j & p_i^T p_k \\ p_j^T p_i & p_j^T p_j & p_j^T p_k \\ p_k^T p_i & p_k^T p_j & p_k^T p_k \end{bmatrix} \qquad (2)$$

where **G** named Gramian matrix is rigid invariant. Because of the symmetry of human face, there are $\mathbf{G}_{11} = \mathbf{G}_{22} = \mathbf{G}_{12} = \mathbf{G}_{21}$ and $\mathbf{G}_{13} = \mathbf{G}_{23} = \mathbf{G}_{31} = \mathbf{G}_{32}$, if we take four points as nose (point 0), mouth (point i), left eye (point j) and right eye (point k). The computing efficiency will be therefore increased significantly.

(2). Local motion model

Δ(p) representing the local motion from tissue deformation are computed from the sets of AUs combination. The 3D motion parameters are estimated from two successive frames of an image sequences. More frame might be used for robust detection.

2.2. Facial expression analysis / synthesis

The analysis/synthesis of facial expression are represented and decomposed by action units (AUs) based on the Facial Action Coding System (FACS) by Ekman and Freisen. The synthesis of facial expression can be carried by the linear weighted AUs combination as shown. A 3-D wireframe face model for model-based face image analysis / synthesis was generated. Its size was extended to about 500 triangle patches for the purpose of improving the quality of synthesized image.

The deformations of 3D wire frame model which cause the facial expressions are carried out such that: In first, the total feature points, which are usually the most expressive point in a face, are to be taken as the first control points. Second, the relation between the first and secondary control points, and between the secondary control points and the vertices of every facial triangle, are built. In third, according the rule of AUs decomposition / combination related to typical facial expressions, various facial expressions could be synthesized. The fourth, the total points with its intensity values placed in each triangle will be moved up on the interpolation from vertices of triangles.

2.3. Textural Mapping and Interpolation

Current technique for synthesizing facial expressions adopt 3-D wireframe as the face model. Textural mapping and interpolation are well used for this purpose. In where the original facial images are regarded as 2-D objects and defined as a finite domain of a plane with gray level associated with each points. The textural mapping for facial image synthesis are on two steps achieved:

(1). The original facial image R(u,v) on system (u,v) are transformed on to the 3-D surface on system (x_s, y_s, z_s).

(2). Make the orthogonal projection from 3-D surface mentioned on to a 2-D screen plane which based on system (x,y).

Now the problem are mainly the $R^2 \rightarrow R^2$ mapping and deal with $R^d \rightarrow R$ interpolation. Let two 2-D points data sets are (x_i, y_i) and (u_i, v_i), mapping are represented by the transform $T = (Tu, Tv)$: $R^2 \rightarrow R^2$:

(1). T are invariant representation in each element transform Tu and Tv.

(2). $T(x_i, y_i) = (u_i, v_i)$, $i=1,2,.....N$.

In our case, the textural mapping of each patch means the transform T from original patch to the deformable ones. The sum of transform of every patch achieve the total textural mapping.

The mapping of a patch from original image to 3-D surface could be

$$(x_s, y_s, z_s) = [u, v, 1] \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{bmatrix} \qquad (3)$$

In case of making orthogonal projection on z axis, there are $x = x_s$, $y = y_s$, and $z = z_s = 0$ i.e. $c_1 = c_2 = c_3 = 0$. Eq.3 will then be degraded to:

$$(x, y) = [u, v] \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \\ a_3 & b_3 \end{bmatrix} \qquad (4)$$

The six coefficients of a_1, a_2, a_3, b_1, b_2, b_3 will be solved by using the vertices coordinates of each patch before and after mapping.

Usually, the mapped points P might not be located on lattices such as the points A, B, C and D shown in Fig .1. Let the four points are

A: (u_A, v_A), g_A C: (u_C, v_C), g_C
B: (u_B, v_B), g_B D: (u_D, v_D), g_D

The gray level g_P of point P can then be derived by dual-linear interpolation:

$$g_p = (1-\Delta u)(1-\Delta v) g_A + \Delta v(1-\Delta v) g_B + \Delta u(1-\Delta v) g_C + \Delta u \Delta v g_D \quad (5)$$

In experiments, a Candide-like 3-D wireframe face model consisted of about 500 triangle patches are created. Fig.6 show the rotating face model controlled by motion estimation.

The synthesized images by using the textural mapping based on Eq.5, as shown in Fig.7, are rather natural.

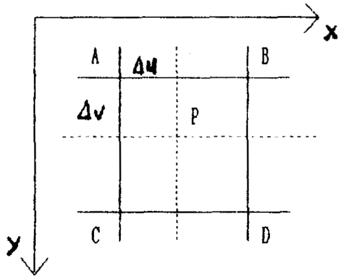


Fig.1 Dual-linear interpolation for textural mapping

2.4. Occlusions in Facial Image Synthesis

The wireframe model of a face in front view may be rotated around 3-D axes and form the 3-D model according with the depth information. The patches of facial wireframe next to the contour in rotated orientation might not be visible after rotation due to occlusion. However, the mapping of occluded patches will be reflected and superimposed on the visible patches next to that contour. The blurs will therefore be taken place as shown in Fig.2. For overcoming this phenomena, the outer normal vector of the patches will be made use to decide which patch have to be hidden. Let three vertices of a triangle patch are A(x_A, y_A, z_A), B(x_B, y_B, z_B) and C(x_C, y_C, z_C) respectively. From Fig.3 the vector AB, and AC will then be

$$\bar{A}\vec{B} = \|X_{AB}\|\vec{i} + \|Y_{AB}\|\vec{j} + \|Z_{AB}\|\vec{k}$$
$$\bar{A}\vec{C} = \|X_{AC}\|\vec{i} + \|Y_{AC}\|\vec{j} + \|Z_{AC}\|\vec{k}$$

where $\|X_{AB}\|$, $\|Y_{AB}\|$, $\|Z_{AB}\|$ are the projection of vector AB on axes of x, y and z. In same case, $\|X_{AC}\|$, $\|Y_{AC}\|$, $\|Z_{AC}\|$ are the projection of vector AC.

The outer normal vector may be introduced as:

$$\vec{N} = \bar{A}\vec{B}\,X\,\bar{A}\vec{C} = \|X_N\|\vec{i} + \|Y_N\|\vec{j} + \|Z_N\|\vec{k}$$

where $\|X_N\|$, $\|Y_N\|$ and $\|Z_N\|$ are the projection of vector N.

$$\|X_N\| = \|Y_{AB}\| * \|Z_{AC}\|$$
$$\|Y_N\| = \|Z_{AB}\| * \|Z_{AC}\| - \|X_{AB}\| * \|Z_{AC}\| - \|Z_{AB}\| * \|Y_{AC}\|$$
$$\|Z_N\| = \|X_{AB}\| * \|Y_{AC}\| - \|Y_{AB}\| * \|X_{AC}\|$$

Now, the viewer space is based on rectangular coordinate system and follow right hand screw rule as shown in Fig 3.

In the case of the objects faced to viewer, the triangle patches have to be hidden and wouldn't be visible if $\|Z_N\| > 0$. Only patches whose $\|Z_N\| < 0$ could be visible. Therefore, the occlusion problem in facial image synthesis could be overcome.

Fig.7 shows the synthesized rotating facial images of "Miss America" in which the blurs are overcome.

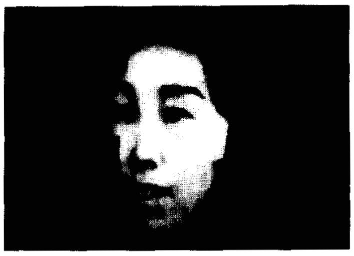


Fig.2 Blurs caused by occlusions in a rotating facial image

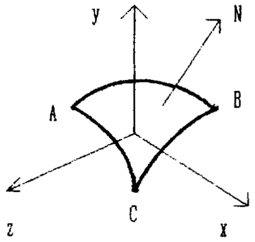


Fig. 3 Triangle patch and its outer normal vector

3. FAST AUTOMATIC FACE FEATURE POINTS EXTRACTION

Real time automatic face feature points extraction for image analysis / synthesis have been the key technique and not been reasonable solved yet in existing papers published. A fast face feature points extraction scheme based on shift template pair for real time applications are proposed. The face feature points belong to eyebrows, eyes, nose and mouth region can be automatically extracted fast with rather high position accuracy.

3.1. Preprocessing

The purpose of preprocessing is to eliminate the undesirable objects on picture frame and separate the human face from its background. We may take an average brightness Γ_f of input image in the one sixteenth central area in the frame. Give a threshold Δ_f and let the value of every pixel in the facial image such that:

$$\Gamma(n) = \begin{cases} \Gamma const \ , & \text{if } \Gamma_f - \Delta_f \le \Gamma(n) \le \Gamma_f + \Delta_f \\ 0 \ , & \text{otherwise} \end{cases}$$

Under some reasonable promises, the face region and contour can be separated by thresholding.

3.2. Template pairs design

We proposed a template pair consisted of two complementary templates for raising extracting correctness. As an example, here we just give the template pairs for extracting feature points in eye region. There are 4 feature points A,B,C and D in eye region (Fig.4) are to be extracted.

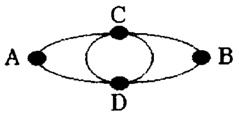


Fig 4 Eye model

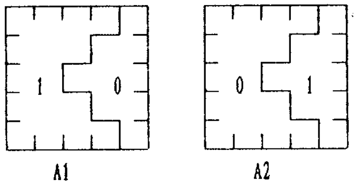


Fig. 5 Template pair for point A

3.3. Extracting skill

Give threshold δ_{A1} for point A and utilize the template A1 to make an exclusive-OR operation with the pixels in eye region. A point set G_{A1} will be resulted if the value from operation are bigger than δ_{A1}. Again, give threshold δ_{A2} and result the points set G_{A2} on same procedure. Make an OR operation for G_{A1} and G_{A2}, we result the point A. There are possible to get a small point set G_A by this OR operation. We may result the point A from averaging G_A.

The feature points B, C, D and other feature points belong to eyebrows, nose and mouth regions may be extracted by similar template pair set on various parameterizing. The total 22 feature points in a face extracted by this skill are achieved.

Experiments for feature points extraction

The test image "American Woman" sized 128x128 will be used for face feature points extraction. The time consumed for extracting total 22 feature points are 0.73 seconds in condition of PC-486/33 with an MS-DOS 5.0 operation system and written in Borland C++3.1 language. The positions of feature points extracted are accurate enough.

It is possible to utilize the method presented in this paper for real time applications (such as video telephone / conference) if the high speed DSP and firmware technique are used.

4. CAUCHY-SCHWARZ-INEQUALITY PERFORMANCE CRITERIA

How to judge the "distortion" or differences between the facial images before and after image analysis / synthesis are still a problem. The criteria used in existing image coding, such as MSE and SNR, are not suitable well for model-based facial image coding. For instance, a little global variations (rotations and translations etc. but no facial expression changes) between two versions of an image will cause significant degradation of MSE and SNR though the face structure and its expressions have no changes. In other hand, the facial expression changes (e.g. eyebrow raise) might not influence the MSE and SNR significantly because the local motions (facial tissue deformations) areas in the face are so small compared to full image frame, though they are meaningful . Therefore, we have to find a new criteria for effective performance judgment in facial image analysis / synthesis. In this paper, an effective criteria based on Cauchy-Schwarz-Inequality (CSI) for facial image coding was presented. As known, the consistency of two images which were referred by two non-zero vectors **u** and **v** are depended on the cosine of the angle between **u** and **v**.

$$CSI\ (u,v) = \cos\left[u,v\right] = u \bullet v/|u||v|$$

It is clear that

(1). CSI(**u**,**v**) <= 1, $\forall$ (**u**,**v**)

(2). CSI(**u**,**v**) = 1, if and only if **u** = k**v**, where k is an non-negative scalar.

As shown, CSI(**u**,**v**) indicate how well **u** corresponds to **v**. In other words, the **u** which maximizes CSI(**u**,**v**) would be the best match to **v**. In our method, the face will segmented into i { i=1,2,...n } regions which are independent each others on expressions (e.g. "eye region", "mouth region" and so forth). The vectors subject therefore to an orthonormal coordinate system in "n" According the projection theorem in n-dimensional Eucliden space, we conducted CSI as follow:

$$CSI_w(u,v) = \left\{\sum_{i=1}^{n} w_i^u w_i^v u_i v_i\right\} \Big/ \left\{\sum_{i=1}^{n}(w_i^u u_i)^2\right\}^{1/2}\left\{\sum_{j=1}^{n}(w_j^v v_j)^2\right\}^{1/2}$$

where w_i^u, w_i^v and w_j^v are weighting factor. In this paper, a new criteria K_w was defined as:

$$K_w = -20\ Log_{10}\left\{1 - CSI_w(u,v)\right\}/CSI_w(u,v)$$

The reconstructed images results from our facial image analysis / synthesis algorithm are gotten based on a deformable 3-D wireframe model. we had computed the CSI$_w$(**u**,**v**), K_w and also the MSE, SNR. The result of experiments are shown that:

(1). The new criteria K_w will be more sensitive than SNR though a slight changes of facial expressions so that quite suitable for facial image analysis / synthesis.

(2). This performance judgment would not be influenced by the global motions of head.

5. CONCLUSION REMARK

Though the model-based image coding have received much attentions for its high compression ratio in very low bit rates condition, there are a lot problems have to be solved before its implementing in practice. In this paper, several key problems such as human face model and its 3-D invariant representation, textural mapping / interpolation, occlusion problem in rotating facial image synthesis, real time automatic face feature points extraction and a new criteria based on Cauchy-Schwarz-Inequality are discussed. In next phase, some problems including the refreshing of reference image based on expression changing and the adaptive facial image synthesis based on by CSI etc. should be further investigated

6. ACKNOWLEDGMENT

This work has been supported by Natural Sciences Foundation of China (NSFC) under grant number of No.69272032.

7. REFERENCES

(1). T.S.Huang, et al. " Motion and Structure from othorgraphic Projection", Trans. Patt. Anal. Mach. Intell., vol.11, No.5, pp 536-540, 1989

(2). A.Pentland, " Linear Shape from Shading ", Int.J.Comput.Vision, vol.4, pp.153-162, 1990

(3). R.Forchheimer,et al, " Image Coding — from Waveform to Anomation ", IEEE Trans on ASSP vol.37, no.12, pp.2008-2032, Dec. 1989

(4). H.G.Musmann, et al. " Object-Oriented Analysis-synthesis Coding of Moving Images ", Signal Processing: Image Communication, vol.1, no.2, pp.117-138, Oct. 1989

(5). D.Terzopoulos, et al., " Analysis and Synthesis of Facial image Sequence using Physical and Anatomical Models ", IEEE Trans. on PAMI, vol.15, no.6, pp.369-379, June 1993

(6). P.Ekman and W.V.Friesen, " Facial Action Coding System ", Consultion Psycologist Press (1977)

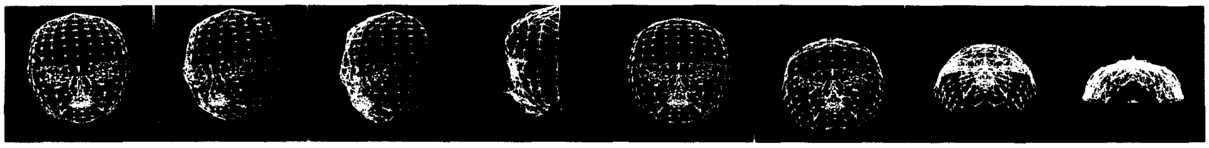
(7). W.J.Welsh, et al., " Model-Based Image Coding ", Telcom. Tech. J, vol.8, no.3, pp.94-105, 1990

(8). T.Kimoto and Y.Yasuda, " Hierarchical Representation of the Motion of A Walk and Motion Reconstruction for Model-Based Image Coding ", Optical Engineering, vol.20, no.7, pp.888-903, 1991

(9). H.Harashima, et al., " Intelligent Image Coding and Communications with Realistic Sensation — Recent Trend ", IEICE Trans on vol. E74, no.6, pp.1582-1592, June 1991

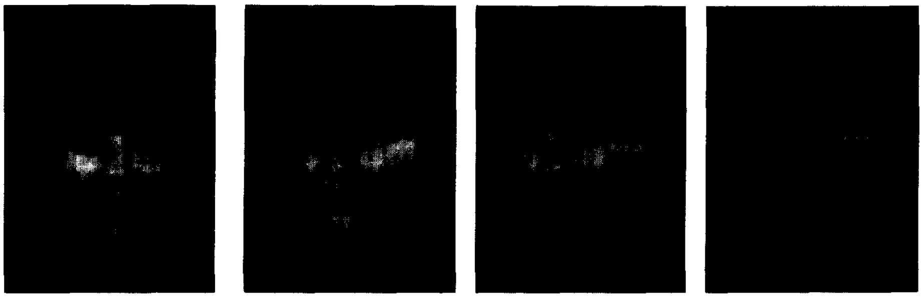
(10). K.Aizawa, et al., " Model-Based Analysis Synthesis Image Coding System for A Person's Face", Signal Processing: Image Communication, vol.1, no.2, pp.139-152, Oct. 1989-114

(11). C.S.Choi, et al.," Analysis and Synthesis of Facial Expression in Knowledge-Based Coding of Facial Image Sequences ", ICASSP'91, pp.2737-2740, 1991



(a) $\phi, \theta, \psi = 0^\circ, 0^\circ, 0^\circ$, (b) $\phi, \theta, \psi = 0^\circ, 30^\circ, 0^\circ$ (c) $\phi, \theta, \psi = 0^\circ, 45^\circ, 0^\circ$, (d) $\phi, \theta, \psi = 0^\circ, 90^\circ, 0^\circ$
(e) $\phi, \theta, \psi = 10^\circ, 0^\circ, 0^\circ$, (f) $\phi, \theta, \psi = 30^\circ, 0^\circ, 0^\circ$, (g) $\phi, \theta, \psi = 60^\circ, 0^\circ, 0^\circ$, (h) $\phi, \theta, \psi = 90^\circ, 0^\circ, 0^\circ$

Fig.6 Rotated 3D face model controlled by MC



(a) $\phi, \theta, \psi = 0^\circ, 10^\circ, 0^\circ$, (b) $\phi, \theta, \psi = 0^\circ, 20^\circ, 0^\circ$, (c) $\phi, \theta, \psi = 0^\circ, 30^\circ, 0^\circ$, (d) $\phi, \theta, \psi = 0^\circ, 90^\circ, 0^\circ$

Fig. 7 Synthesized images "Miss America" after textural mapping and blur overcoming