Automatic 3D face synthesis using single 2D video frame

Y. Sheng, A.H. Sadka and A.M. Kondoz

3D face synthesis has been extensively used in many applications over the last decade. Although many methods have been reported, automatic 3D face synthesis from a single video frame still remains unsolved. An automatic 3D face synthesis algorithm is proposed, which resolves a number of existing bottlenecks.

Introduction: Automatic 3D face synthesis plays a crucial role in many applications. One example is real-time 3D model-based video coding. Here, a generic 3D face model should ideally be adapted automatically to a human face in the first video frame. Additionally, facial texture should be acquired without supervision. However, the time-consuming 3D scanner [1] is not accepted for such an automatic system. The method of using two orthogonal photos [2] cannot provide real-time processing either. Feng [3] synthesised the face only from a single image, but this method needs to estimate head rotation parameters by using another reference image. Moreover, it hardly brings accurate results in 3D face model adaptation by only using three facial features. Strictly speaking, there is no fully automatic 3D face synthesis from a single image proposed to make such a real-time video application a reality. This Letter presents an automatic synthesis system, enabling 3D face modelling from an arbitrary 2D header-and-shoulder video frame without partial occlusion.

Facial feature extraction: Automatic facial feature extraction comprises two phases: face detection and facial feature extraction. As a fast and generic clue, the distribution of skin tone in chrominance Cr is first utilised to segment the face, followed by a statistical method of calculating standard deviations of 4 × 4 pixel blocks for hair removal, because luminance change in the face region is less evident than in that of the hair [4]. Facial features serve as references for adapting the generic face model. Too many features would increase computational complexity, while too few features would result in inaccurate fitting of the face model. Fig. 1 shows some selected features, with the corresponding extraction approaches used in our system (more detail can be found in [5]). These features topologically form the basis of vertices of the 3D generic face model. They provide geometrical information that helps in adapting the 3D face model onto a specific individual face in the 2D video frame.

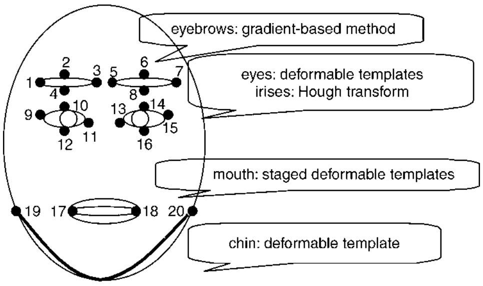
eyebrows: gradient-based method
eyes: deformable templates
irises: Hough transform
mouth: staged deformable templates
chin: deformable template
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Fig. 1 *Facial feature extraction illustration*

Face model adaptation: This is a process in which the generic 3D face model is deformed to fit a specific face. This process can be broken down into two steps, global adaptation and local refinement. In our system, Candide-3 shown in Fig. 2 is chosen as the face model because it is easy to use as well as compliant with MPEG-4 [6].

1) Global adaptation: The rigid motion from an arbitrary point $P(X, Y, Z)$ on the surface of a 3D face model to the destination $P'(X', Y', Z')$ with infinitesimal Eulerian angles can be formulated as:

$$\begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix} = \begin{bmatrix} 1 & -\Delta\theta_Z & \Delta\theta_Y \\ \Delta\theta_Z & 1 & -\Delta\theta_X \\ -\Delta\theta_Y & \Delta\theta_X & 1 \end{bmatrix} \begin{bmatrix} S_X & 0 & 0 \\ 0 & S_Y & 0 \\ 0 & 0 & S_Z \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + \begin{bmatrix} T_X \\ T_Y \\ T_Z \end{bmatrix} \tag{1}$$

where θ_X, θ_Y and θ_Z are rotation angles around 3D axes X, Y and Z; S_X, S_Y and S_Z are scale factors; and $(T_X, T_Y, T_Z)^T$ is the translation vector. Global adaptation can equivalently be thought of as estimation of these nine rigid motion parameters. Assuming orthographic projection, the translation vector can be derived by calculating the distance from the 3D face model to the 2D face centre. Let P_l, P_r, P_c and P_m, respectively, indicate the centre of the left eye, the centre of the right eye, the middle point between two eyes and the mouth centre of the planar face, while P_l', P_r', P_c' and P_m' are the corresponding features in the 2D projection of the 3D face model. So the scale factor S_X and S_Y are defined as $S_X = \|P_l - P_r\| / \|P_l' - P_r'\|$ and $S_Y = \|P_c - P_m\| / \|P_c' - P_m'\|$. Since the depth of the face is invisible on the planar frame, for the sake of automation, S_Z is evaluated by simply averaging S_X and S_Y due to linear dilatability. The inclination of the head, i.e. θ_Z, is obtained by measuring the angle between line segment P_lP_r and the 2D horizontal axis. A cross-section-oriented method reported in [7] is employed to find θ_Y. As for θ_X, thanks to the deficiency of depth information, it is presumed that the head is neutral with respect to axis X, that is $\theta_X = 0$.

2) Local refinement: Four facial features of the face model, i.e. eyebrows, chin, eyes and mouth are in turn refined in our system. For eyebrow refinement, eight eyebrow vertices (black-spotted in Fig. 2) are simply translated to fit eight extracted eyebrow feature points (Fig. 1). In chin refinement, five chin vertices of the 3D face model (black-spotted in Fig. 2) are stretched or shrunk towards the extracted 2D chin contour. With the study of anthropometry and muscle motion, three vertices in the middle are translated along the line segments formed with the upper black-squared vertex, while the other two vertices are lined up with two black squares diagonally above them. Mouth refinement consists of affine transformation and contour adjustment. All the mouth vertices embraced within a rectangle in Fig. 2, are globally 2D affine transformed by calculating the difference between the extracted planar mouth and the 2D mouth projection of the globally adapted 3D face model, in terms of translation, scaling and rotation. However, the affine transform cannot guarantee matching of the lip contours. To retain the natural shape of the outer upper lip contour, all the vertices on the contour are translated towards the extracted 2D contour with a constant, estimated by evaluating the distance between the middle point of the outer upper lip contour of the planar face and that of the 2D projection of the 3D face model. The inner upper lip contour is adjusted by simply translating the vertices to intersect the extracted counterpart. Similarly, the lower lips can be fitted using the same means. The principle of eye refinement is identical to that of mouth refinement. The eyelids of the face model are adjusted towards the extracted eye contours, following an affine transformation of the eye vertices of the face model, embraced within two ellipses. However, the eyes must be treated as two separate objects, using two separate sets of parameters. In addition, linear interpolation is applied for adapting those non-boundary vertices.

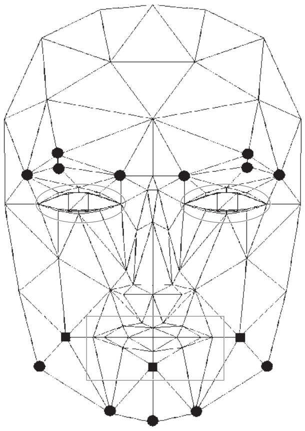

Fig. 2 *Candide-3 face model*

Face synthesis: Generally speaking, face model adaptation produces three outcomes, i.e. a personalised texture map, corresponding texture co-ordinates of the face model, as well as a set of individualised 3D vertex co-ordinates. In our system, a 256 × 256 pixel texture map, in which the face is centred, is automatically acquired from a CIF video frame. For the sake of processing convenience, the face must appear as centred as possible in the texture map. In fact, the position of the face is hardly constantly centred in the video frame because the speaker randomly sways. So the location of the face in the video frame is studied. We segment both the horizontal and vertical axes into three intervals with the landmarks $L_T/2$ and $(L_X - L_T/2)$ horizontally, and $L_T/2$ and $(L_Y - L_T/2)$ vertically, where $L_T/2$ indicates half a width of the texture map (128 in our system), L_X and $L_Y$are the width and height of a CIF video frame, respectively. These four landmarks decompose the video frame into nine districts. According to the location of the face centre (C_Y, C_X) with respect to these districts, we can decide how to create the texture map. For example, if $C_X < L_T/2$ and $C_Y < L_T/2$, then the co-ordinate $(L_T/2, L_T/2)$ will be regarded as the centre of the texture map, while if $L_T/2 < C_X < L_X - L_T/2$ and $L_T/2 < C_Y < L_Y - L_T/2$, then the texture map will be centred at (C_Y, C_X). Based on the automatically generated texture map, texture co-ordinates of the 3D face model can be generated by mapping the 3D co-ordinates into the texture space. With all the above and the previously individualised 3D vertex co-ordinates in face model adaptation, a specific 3D face model is finally synthesised.

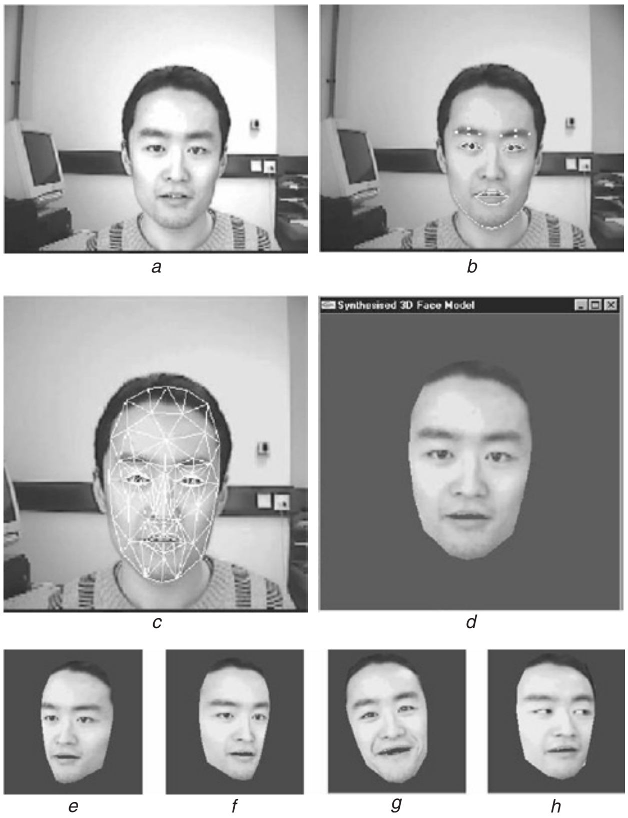

Fig. 3 *Experimental results*

Algorithm illustration:
a Head-and-shoulder 2D video frame
b Extracted facial features
c Adapted Candide-3 wireframe
d Synthesised 3D face
e and *f* Rotated 3D faces
g and *h* Other synthesis results

Results: 156 video frames from different video sequences with complex backgrounds in CIF format were tested, 92.3% of which can be successfully synthesised using the proposed algorithm. Unlike the existing algorithms, our method models 3D faces only from a single frame without supervision. The test results show that the proposed system is capable of robustly coping with the 2D head-and-shoulder frames in different expressions, head rotations and race variability. Figs. 3*a*, *b*, *c* and *d* illustrate outputs from each stage of the algorithm. Figs. 3*e* and *f* are the rotated views of the synthesised 3D face in image *d*, while Figs. 3*g* and *h* show synthesised 3D faces from different 2D images. Moreover, averaging PSNR of the synthesised 3D faces shows that the system can achieve a mean of 38.7 dB in comparison with the original 2D faces, while it is only 35.4 dB using the adaptation method proposed in [3] for the same frames.

Conclusions: An automatic algorithm is reported, which enables 3D specific face synthesis from an arbitrary head-and-shoulder video frame with a complex background, by subtly overcoming the ill-posed problem of estimating 3D information from a 2D image.

© IEE 2004 *4 June 2004*
Electronics Letters online no: 20045698
doi: 10.1049/el:20045698

Y. Sheng, A.H. Sadka and A.M. Kondoz (*Centre for Communication Systems Research, University of Surrey, Guildford GU2 7XH, United Kingdom*)

References

1 Eisert, P., and Girod, B.: 'Analyzing facial expressions for visual conferencing', *IEEE Comput. Graph. Appl.*, 1998, **18**, (5), pp. 70–78

2 Goto, T., Lee, W., and Thalmann, N.: 'Facial feature extraction for quick 3D face modeling', *Signal Process., Image Commun.*, 2002, **17**, pp. 243–259

3 Feng, G.C., and Yuen, P.C.: 'Recognition of head-&-shoulder face image using virtual frontal-view image', *IEEE Trans. System Man Cybern. B, Cybern.*, 2000, **30**, (6), pp. 871–883

4 Sheng, Y., Sadka, A.H., and Kondoz, A.M.: 'Automatic face segmentation for 3D model-based video coding'. IEE Int. Conf. on Visual Information Engineering, UK, July 2003, pp. 274–277

5 Sheng, Y., Sadka, A.H., and Kondoz, A.M.: 'An automatic algorithm for facial feature extraction in video applications'. 5th Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'2004), Lisbon, Portugal, April 2004

6 Ahlberg, J.: 'Candide-3: an updated parameterised face'. Report No. LiTH-ISY-R-2326, 2001 (Linkoping University, Sweden)

7 Reinders, M.J.T., and Lubbe, J.C.A.: 'Facial feature localization and adaptation of a generic face model for model-based coding', *Signal Process. Image Commun.*, 1995, **7**, pp. 57–74