MVA2000 IAPR Workshop on Machine Vision Applications, Nov. 28–30, 2000, The University of Tokyo, Japan

8—32

Human Face Image Creation for Virtual Space Teleconferencing using Camera Input Images

Hajime Sato*
Graduate School of Global Information and Telecommunication Studies, Waseda University

Nobuyoshi Terashima† Hideyoshi Tominaga‡
Global Information and Telecommunication Institute, Waseda University

Abstract

Since the human face plays an important role in man-to-man communication, systems enabling distant users to communicate while viewing each others' face have been demanded. Recently, virtual teleconferencing systems which bring distant users together through the communication network by displaying avatars in three-dimensional virtual space have been developped to satisfy these demands. A high-speed and high-precision human face image creation method using camera input images is proposed to realize natural communication in such systems. A fundamental experiment using a single person's face showed good results that realistic human face images from arbitrary directions could be created and displayed with high speed and precision.

1 Introduction

The human face plays a very important role in everyday man-to-man communication. A great deal of information could be obtained naturally from motions such as blinking, gaze directions, facial expressions and gestures. Hence, the development of systems which enable users in distant places to communicate while viewing each others' face have been highly demanded.

Videophone and Teleconferencing Systems have been created hitherto, but had drawbacks such as the difficulty of acquiring eye contact, and lack of the sense of closeness between users. In recent years, Virtual Teleconferencing Systems have been developped to make communication in teleconferencing closer to communication in real life[1]. These systems enable users to interact by displaying avatars in computer generated three-dimensional virtual space.

If realistic human face images from arbitrary views could be created and displayed with high speed and precision, a higher sense of closeness between distant users in such systems could be expected. We propose a human face creation method using camera input images, which could be implemented using an ordinary video camera and computer.

*Address: 1-3-10 Nishi-waseda, 29-7 Bldg., Shinjuku-ku, Tokyo 169-0051 Japan. E-mail: hajime@giti.waseda.ac.jp
†Address: 1-3-10 Nishi-waseda, 29-7 Bldg., Shinjuku-ku, Tokyo 169-0051 Japan.E-mail: terasima@giti.waseda.ac.jp
‡E-mail: tominaga@giti.waseda.ac.jp

2 Concept

In conventional techniques based on the concept of Model-based Coding[2], facial movements were extracted and analyzed from two-dimensional face images, and reflected upon three-dimensional expressionless face models created beforehand. However, it is difficult to extract and parameterize three-dimensional face feature movements accurately, and therefore complicated to reconstruct natural- looking face images. Systems using optical markers and magnetic sensors to catch real time facial movements have also been constructed[3], but it is troublesome to mount devices on the exact positions, and it is also uncomfortable for the participant.

In virtual teleconferencing, accurate parameterization of the human face is not neccessarily essential, and it is sufficient if an object maintaining the features of a two-dimensional camera input image could be created and displayed at real time[4]. Our proposed method is based on this concept and is carried out as follows.

[Step 1] The human face region is extracted and modified using a three-frame difference between consecutively input frames.

[Step 2] The face image acquired in **Step 1** is matched with a set of three-dimensional templates. By using a template interpolation matching method, the human face direction in horizontal, vertical, and diagonal is calculated with high speed and accuracy .

[Step 3] The face position and direction parameters obtained in **Step 1** and **Step 2** are reflected upon a simplified three-dimensional model of the human face created beforehand. Camera input images are texture mapped directly onto this model to reproduce photorealistic face features and expressions.

Figure 1 shows the flow of the steps mentioned above. The details of each step are described in the following sections 3 to 5.

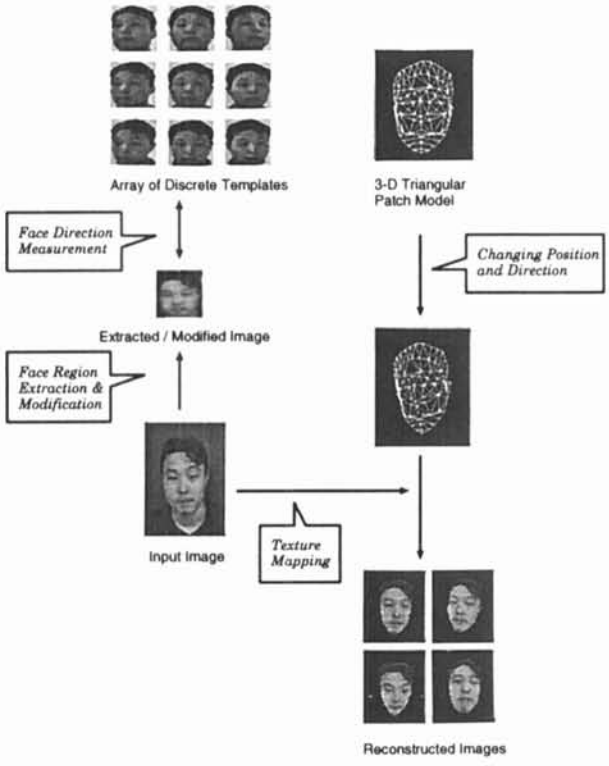

Figure 1: Flow of Proposed Method

3 Face Region Extraction and Modification

As shown in Figure 2, the face region is extracted from video input frames using a three-frame difference method[5].

The binary differences between the present frame f_i, and the previous and following frames f_{i-1}, f_{i+1} are calculated. By calculating the AND of these two black-and-white differential images, the position of the face region in frame f_i could be estimated from the distribution of white pixels. The face region is determined by using a horizontal and vertical histogram and clipped out. The vertices coordinates of determined face region is used later in Section 5.

The size of the extracted face region image is modified to n x n [pixels], and to reduce effects of noise and changes in facial expression, the image is smoothened with a filter.

4 Measurement of Face Direction

4.1 Template Matching

A set of three-dimensional templates is created beforehand, by rotating the camera horizontally,

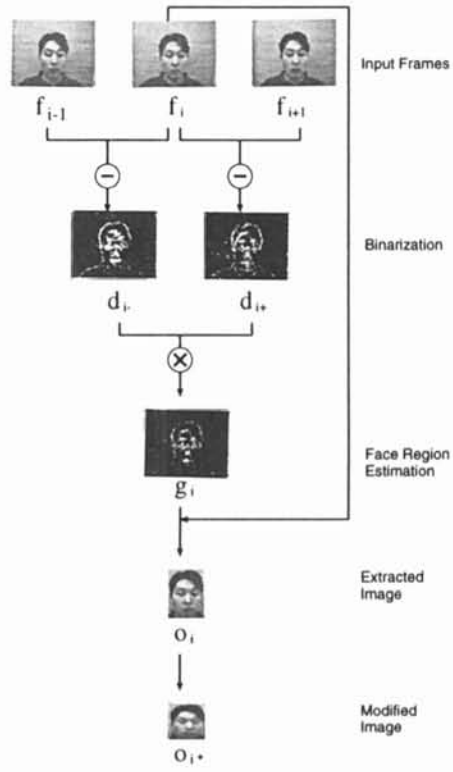

Figure 2: Human Face Region Extraction and Modification

vertically, and diagonally at even intervals, to acquire expressionless human face images from various angles. The extracted and modified face image o_i* obtained in section 3 is matched with the templates using a simple similarity method[6].

4.2 Interpolation

A conventional template interpolation method proposed by Koike and Tanabe[6] is extensively used so the face direction parameters could be estimated with high accuracy even when there is no template which matches perfectly with the input image.

The best fit block of eight neighboring templates is selected using the template matching method mentioned earlier in Section 4.1. As shown in Figure 3, we assume that the center of imaginary template f_t, which matches perfectly with the input image, exists inside a cube created by connecting the centers of the eight neighboring templates f_1 to f_8 . The center of f_t divides the height, width, and depth of the cube by $p : (1-p)$, $q : (1-q)$, $r : (1-r)$ respectively.

Let m_1, m_2, m_3, m_5 be the simple similarities of templates f_1, f_2, f_3, f_5 respectively. If changes in similarity between two neighboring templates are assumed linear, p, q and r are represented as follows[7].

$$p = \frac{1 - m_1}{2 - m_1 - m_2}$$

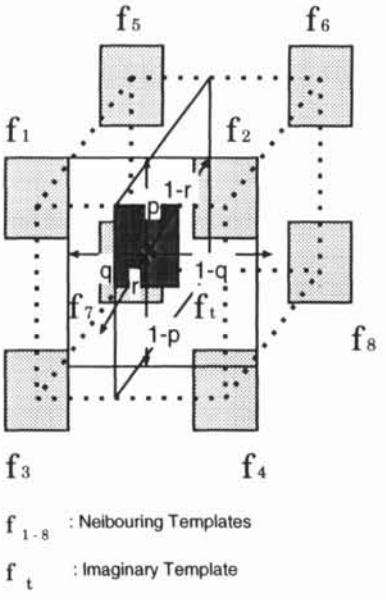

Figure 3: Template Interpolation

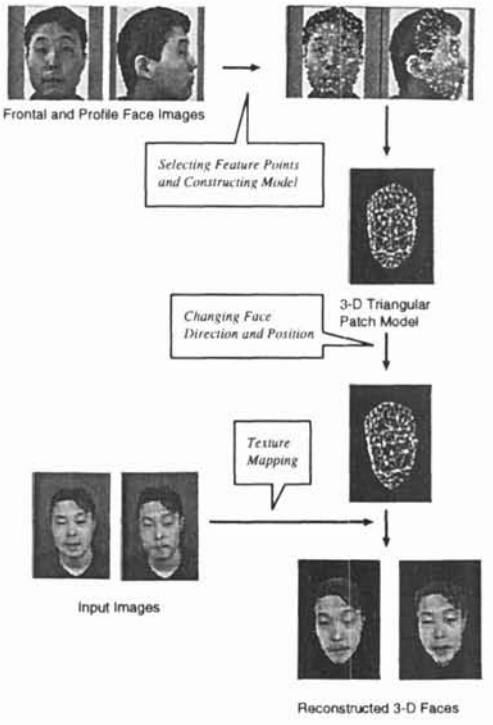

Figure 4: Three-dimensional Face Image Construction

$$q = \frac{1 - m_1}{2 - m_1 - m_3} \tag{3}$$

$$r = \frac{1 - m_1}{2 - m_1 - m_5}$$

5 Three Dimensional Face Image Reconstruction

A frontal and profile image of the same person's face is input into a designated editor. Feature points, such as the edges of facial components are specified on the frontal image. Its corresponding points are specified on the profile image, and the three-dimensional coordinates of each feature point is obtained. By connecting these coordinates, a three-dimensional triangular patch model of the human face is created[8].

This model is moved and rotated according to the position and direction parameters obtained in sections 3 and 4. Each triangular patch is judged if it is facing obverse or reverse. For each triangular patch facing obverse, the two-dimensional texture is extracted from the input image, and modified using an affine transform according to the coordinates of the feature points.

Changes in facial expression could be approximated by simply remapping the texture if we ignore the movements of the chin and upper cheek region. The vertices of the triangular patch model are not altered during the process, and facial expression changes are reproduced by mapping the input images consecutively in sequence.

6 Experiment

6.1 Implementation

To confirm effectiveness of our proposed method, a fundamental experiment using a single person's data was conducted on a UNIX workstation (MIPS R4400 200MHz) under the following conditions.

Input Images Images of a single person talking naturally at the camera were obtained under indoor lighting and a plain background using a home-use video camera. Frames were extracted at 10 [f/s], and modified into 8bit grayscale RGB images of 320 x 240 [pixels] size.

Templates set The participant's face was scanned using a three-dimensional laser range scanner, and an ideal three-dimensional human face was rendered. A 5 by 5 by 5 template set (consisting of 125 templates) of face images obtained at an interval of 15[deg] in a range of +/-30 [deg] in horizontal, vertical and diagonalwas created. All templates were 8bit grayscale RGB images of 64 x 64 [pixels] size.

3-D Model A three-dimensional triangular patch model consisting of 83 feature points and 129 patches was created. More points were specified where there were complex contours.

Since the modules of our proposed method were developped independently, the entire process was not completely automated.

6.2 Results and Discussion

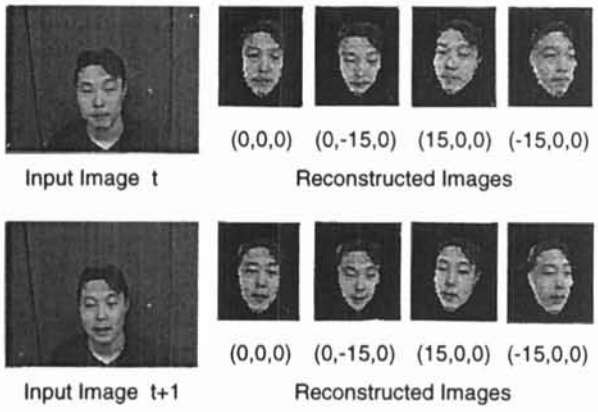

Figure 5: Experimental Results (1)

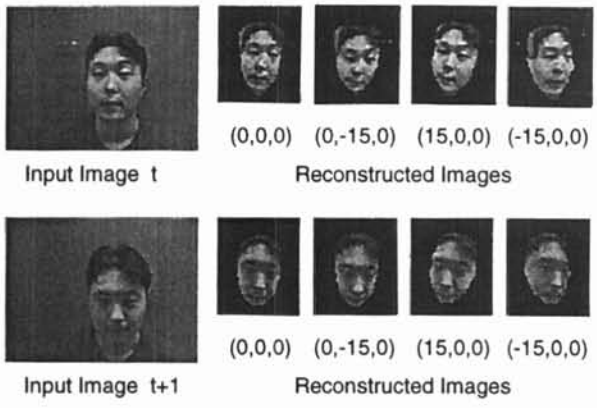

Figure 6: Experimental Results (2)

Experimental Results are shown in Figures 5 and 6. (x, y, z) indicates that the face direction is rotated x degrees horizontally, y degrees vertically, and z degrees diagonally.

Figures 5 and 6 shows a good result that face images from arbitrary directions could be reconstructed with high precision. Also, by comparing the reconstructed images of Input Image t with those of Input Image t+1, it is noticable that changes in facial expressions could be reproduced by changing the model's texture.

As shown in Figure 6, distortion was detected from the reconstructed face images when there was a large movement in face direction. This is due to errors in face direction measurement, and the texture wasn't mapped correctly. It is neccessary to investigate both our direction measurement algorithm, and compensation between the input image and template. Also, the reconstructed face images tend to become unnatural when there is a large change in the face model's direction. This should be improved by increasing the number of cameras and contriving camera positions.

7 Conclusion

We proposed a high-speed and high-precision method using camera input images to reconstruct human face images from arbitrary directions as a basic technology for virtual space teleconferencing. Also, our experiment using a single person's data showed good results, and proved that our method could be implemented using an ordinary video camera and computer. Future work include the improvement of individual techniques as mentioned in section 6.2, and the construction and evaluation of a practical system.

References

[1] F. Kishino, "Human Communication" (in Japanese), Journal of the Institute of Television Engineers of Japan, Vol. 46, No. 6, pp 698-702 (1992)

[2] M. Kaneko, Y. Hatori and A. Koike, "Coding of Facial Images Based on 3-D Model of Head and Analysis of Shape Changes in Input Image Sequence" (in Japanese), Trans. of Institute of Electronics, Information and Communications Engineers, Vol.J-71-B, No.12, pp. 1554-1563 (1988)

[3] J. Ohya, Y. Kitamura, F. Kishino, N. Terashima, H. Takemura, and H. Ishii, "Virtual Space Teleconferencing: Real-Time Reproduction of 3D Human Images", Journal of Visual Communications and Image Representation, Vol. 6, No. 1, pp 1-25 (1995)

[4] Y. Mukaigawa, Y. Nakagawa and Y. Ohta, "Face Synthesis with Arbitrary Pose and Expression from Several Images -An integration of Image-based and Model-based approach-", Proc. of ACCV '98 , pp. 680-687 (1998)

[5] T. Agui and T. Nagao, "Computer Image Processing and Recognition" (in Japanese), Shokodo, pp. 162-3 (1996)

[6] Y. Koike and M. Tanabe, "A Pattern Matching Method for Gray-Scale Images Using Template Interpolation and Neural Networks by Weighed Learning" (in Japanese), Trans. of Institute of Electronics,Information and Communications Engineers, Vol.J-75-D-II, No.7, pp. 1151-1159 (1992)

[7] H. Sato, N. Terashima and H. Tominaga, "Human Face Pose Estimation for Creating Avatars " (in Japanese), Technical Report of the IEICE, PRMU99-153 (1999)

[8] Y. Ikezaki, H. Sato, N. Tsuda, N. Terashima and H. Tominaga, "Real-time synthesis of 3-D human face using two cameras " (in Japanese), Proc. of IEICE Spring Conference, D-12-120 (1998)