〔非公開〕

TR－C－0030

Facial Image Processing and Face Modeling

徐　剛　　永嶋　美雄　　小林　幸雄

XU GANG　　YOSHIO NAGASHIMA　　YUKIO KOBAYASHI

1989. 4. 4

ＡＴＲ通信システム研究所

FACIAL IMAGE PROCESSING AND FACE MODELING

Gang Xu, Yoshio Nagashima and Yukio Kobayashi

Artificial Intelligence Department

ATR Communication Systems Research Laboratories

Sanpeidani, Inuidani, Seika-cho, Soraku-gun

Kyoto 619-02, Japan

e-mail address: xu@atr-sw.atr.junet

Abstract

The main goal of this research is to generate three-dimensional facial models from facial images and to synthesize images of the models virtually viewed from different angles, which is an integral component of the ATR virtual space conferencing system project. In this technical report we propose a stereo-based approach to facial modeling, which is currently being implemented. The first part of this report presents a detailed survey of previous research efforts to develop techniques of facial images processing, facial modeling and facial animation. The second part describes the general approach and specific problems and solutions.

The originality and significance of this work lie in that the system can generate a face model without a human operator's interaction with the system.

1. Introduction

The focus of research activities in the AI department is the virtual space conferencing system project, which was initiated in the first half of 1988. The project is very unique in that it can be completed only by collecting people and technologies from various disciplines, such as artificial intelligence, telecommunications, computer vision, human interface, computer graphics, perceptual psycology, and *etc.*

The general aim of the project is to achieve greatest realism of visual presence in teleconferencing. The system is designed to first generate a full 3D model for each participant from his/her images, then to put them into a virtual conference room, and finally to send back stereo images synthesized in consistency with the assumed spatial relations among the participants in the conference room.

The work described in this paper accounts to the first part and the third part, i.e., to model a human face from its two images and to synthesize facial images virtually viewed from different angles.

Communication is one of the basic needs for human beings. One certain way is to meet and exchange words. If the distance is too long, people used to write letters, but it takes long to receive responses. One alternative is to send and receive signals and meaning through electronic or optical channels. Telephone is a great invention to mankind, and its influence over mankind's life style is tremendous. For the generations who were born after telephone sets entered ordinary people's homes, they use it to talk just as they use sticks to eat. One advantage is that talking by telephone people feel free; you can put your feet on the desk, you can be in pajamas, you do not have to care how you look. However, the same advantage can become disadvantage. What telephone can carry is verbal information. But sometimes people need a sort of confrontation, to look into the opposite's eyes, to catch every possible reflection of the mind on the face. Thus what we need is not only verbal information but also visual information. The videotelephone and various teleconferencing systems are invented to meet this need. One problem of these systems is that the monitor

that you are looking at and the camera that is looking at you are not at the same location. Another problem is that one is always aware of the distance between them. The physical distance always brings mental distance, causing difficulties in friendly exchanges of opinion and experience. To overcome these problems in current visual communication systems is exactly what we are pursuing in the project.

Of the visual information in communication, face plays the most significant role. The diversity of face forms in terms of age, sex and race is enormous. It is these forms that allow us to recognize individuals. Even for the same person, face form varies considerably with expression. It, together with gesture, provides complex non-verbal signals which, usually function as an aid to verbal communication, but sometimes can be the main mode of communication, as when one is in a foreign country whose language he does not speak, as when lovers' staring at each other is more valuable than any words, and as when two lines of tears fall down on the face.

During the past two decades, there has been an increasing interest in the information processing of human faces. There are generally three fields in which the research is active: coding for communication, computer graphics and image processing and understanding. A detailed survey is to be given in Section 2. Our general impression of the work in the past years is that one has to make full use of the knowledge about human face if the processing is to be efficient, whichever the field is.

True it is in our case, too. We will use the knowledge two times. One is essentially the two-dimensional relative positions of face features such as eyes, mouth and nose, used as a first-place estimate to extract them from a front view or a side view. The other is a three-dimensional base face model represented by a triangular polyhedrons, which is modified to produce a face model for the specific person according to the 3D positions of the face boundary and features acquired in the earlier stage. The details will be respectively presented in Section 4 and Section 5.

Section 3 gives a general discussion on the strategy taken in the approach. Section 6

describes the synthesis of images virtually viewed from an assumed angle given the 3D face model at hand. Image intensity values are determined by mapping them to the pixels in either the front view or the side view. The last chapter, Section 7 summarizes the previous sections and gives concluding remarks.

Part One

2. A Survey of Previous Research

The papers surveyed below, except a survey article, are classified into three groups: *facial image processing*, *facial modeling* and *facial animaton*. A summary is attached to each paper, given in the form of the first author's name followed by the paper title.

Harashima, *Recent trends in analysis/synthesis coding systems for facial images.* This article is a survey paper. The summary is as follows: The emerge of intelligent coding or knowledge-based coding has added a new perspective to picture coding. It is also called the 5th generation coding. In the case of facial images, the ideal coding system can extract necessary information on shape change and send it to the receiver. The receiver, sharing the base face model with the sender, modifies the base model according to the received shape change information.

the paper group of facial image processing:

The problem of facial image processing is that image intensity alone does not guarantee

a clear segmentation of the image into meaningful regions. Thus it is necessary to make use of the knowledge about the face structure. This is also the central theme of the papers introduced in this group.

Doyama et.al., *Feature extraction for automatic identification of facial images.* Features are separated from face boundary by use of the knowledge that they are always long in the horizontal direction.

Sakai, et. al., *Computer analysis of photographs of human faces.* This paper describes the extraction of face features such as eyes, nose, mouth and cheek by making use of the knowledge about face structure. The characteristic of this system is the stability and reliabiliy of the system's performance.

Seki & Hashimoto, *Feature points extraction for the human face picture.* This paper describes a multi-leveled template-matching approach to facial feature extraction. The first one is a 2nd-order differential, the second one a narrow-long mountain-shaped filter, and the third one a multi-window mask synthesized according to the structure of face features

Yang, et.al., *Model-based approximation of profile edge in human face recognition.* Employing a profile edge approximating the extremal boundaries generated from a face model, the difference between it and the image data is computed and fed back to modify the profile to reduce the difference. After several iterations, a complete profile edge is obtained from the input image with noise.

the paper group of facial modeling:

The central problem in facial modeling is what kind of model to use and how to specify the parameters in the model based on the information available from the input image(s).

Aizawa et.al., *Modeling a person's face and synthesis of facial expressions for use in a model-based synthesis image coding system.* A triangular polyhedron model is generated by affine-transforming a base model to fit the input front view. The are two approaches

to facial expression synthesis. One is to modify the regions in the face model that change shape frequently. The other is to send image patches for those regions that are expressive and relatively stable.

Akimoto, et.al., *face model synthesis from front/side views and 3D base model.* A base model and a front and a side views are given. An operator inputs the position data of the feature points and the face boundary with the help of the front and side views of the face, and modifies the base model according to the position data.

Muragami, et.al., *A study on image generation and transformation for human faces.* A hierarchical representation of face model with triangle units is proposed. Shape modification is basically linear once the translations of some face surface points are given.

Numazaki, et.al., *Model-based identification of persons from facial images* This paper proposes a generalized cylinder model with ellipse cross-sections to represent head. The major axis and the minor axis for each cross-section are determined from the front view and the side view, respectively. Once the model is obtained, images are synthesized and matched with input images to identify persons.

Noguchi et.al., *A simple modeling method for a facial shape.* Markers are fixed on the face, and a front and a side views are taken. The markers are identified and matched among the images. The 3D coordinates of the matched markers are calculated to produce a triangular ploy hedron face model by Delaunay triangulation.

Parke, *Parameterized models for facial animation.* A general discussion of parameterization of face models and a specific parameterized model for facial animation are described.

the paper group of facial animation:

The central problem in this group of papers is how to find a minimum set of features to represent the face shape change and how to modify the base model given the set of

motion vectors.

Akimoto, et.al., *Expressive facial image generation by automatic shape modification.* Based on FACS (the Facial Action Coding System), the jaw rotation and the translation of a few (15) feature points are selected to represent facial shape change. The motion vectors of the non-feature vertices in the triangular polyhedron model are determined by the jaw rotation and the translation of the feature points.

Kaneko, et.al., *Coding of face images based on 3D model of head and analysis of shape changes in input image sequence.* Motions of the mouth, jaw and eyes are estimated and the corresponding points in the model are modified.

Nemoto, et.al., *Synthesis system of 3D facial animation.* A face is divided into 26 components and each is interactively synthesized and modified to achieve smooth facial animation.

Waters, *A muscle model for animating three-dimensional facial expression.* A model of the muscle process is introduced to facilitate more natural facial animation.

Part Two

3. General Strategy

What we want is a 3D model for a face, and what we have at hand is two 2D images of that face. A great gap exists between the two ends. There are three possible ways to narrow the gap: to pull back the stop end, to push forward the start end and to fill something in between.

Let us first look at the stop end. A 3D model can be described in different representations [Marr & Nishihara, 81]. For a face model, there are two choices. One is the generalized cylinder, with cross sections being ellipses and the nose being approximated by small triangles attached to the ellipses [Numazaki, 87] (Figure 1). For each cross section, the only thing to do is to determine the major and minor axes, and if it crosses the nose, the height of the nose at that cross section. The other choice is the (triangular) polygonal representation [Akimoto, 88; Noguchi, 88; Parke, 82] (Figure 2). It has two layers of information, the topology and the vertex positions. To describe a face, the number of vertices is usually above 500. Comparing the two, one finds that the number of parameters to be specified in the second representation is greater than that in the first choice. At the same time, the descriptive power of the second one is much greater than that of the first one, especially when local sharp shape changes are to be described. The difficulty with the second representation lies at the inability to determine the positions of all vertices directly.

There is a trade-off between the descriptive power and the construction difficulty of the representations. One would select the generalized cylinder representation if its descriptive power is enough for the specific task at hand. Examining carefully the sufficiency of the representations, we find that a triangular polygon model is much more desirable, if not necessary, in our case, because a generalized cylinder with ellipse cross sections is too rigid to reflect the minute shape changes of the eyes and mouth.

Now let us turn to the start end. Since we take the stereo-based approach, the

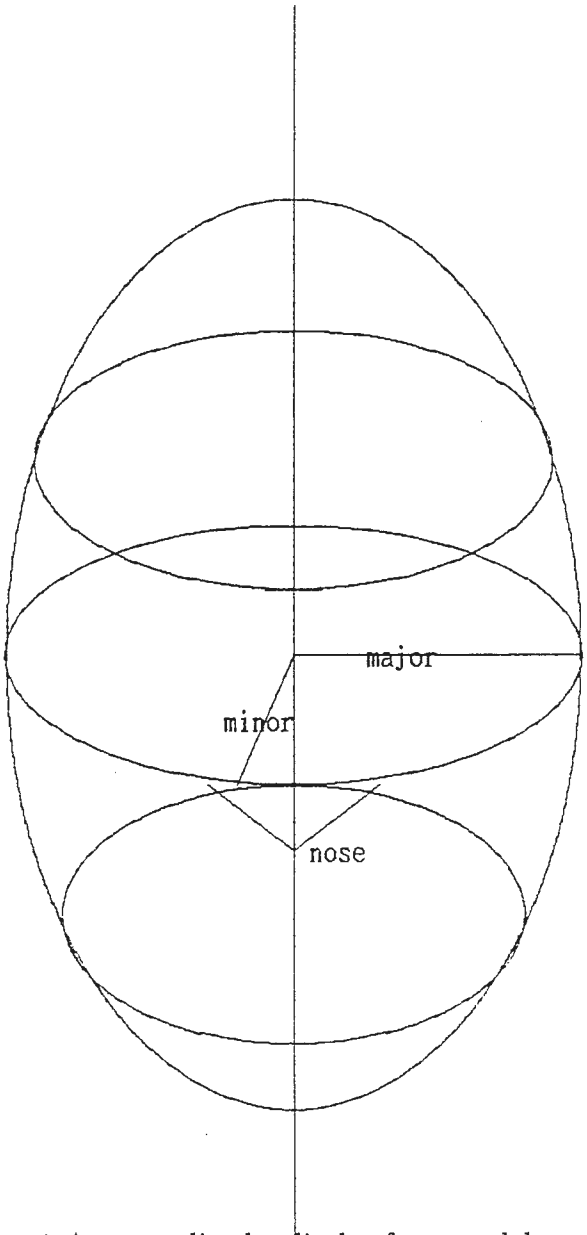

Figure 1 A generalized cylinder face model

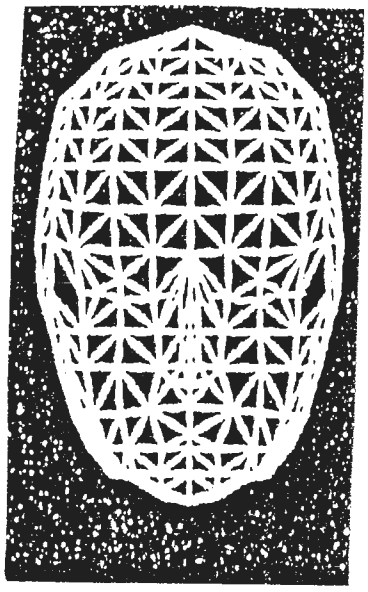

Figure 2 A triangular polygon face model

restriction is that not all face points, but only feature points that are extracted in both images can be matched and can be assigned 3D data. In other words, if a feature is not extracted, then its 3D position cannot be computed. Thus the first step is how to extract face features robustly. However, one has to admit that whatever filters you use, no perfect edge images exist. Even if a perfect edge image does exist, 3D positions of all face points are not available.

Evidently, the stop end of a full 3D model and the start end of only partial 3D data do not match each other, if something is not filled in between. The idea forced by the necessity of an intermediary is that of a base model. That is, a base face model is first built of a certain person, and it is then modified according to the 3D data acquired by matching the front and side views of the face in question.

A minimum requirement for the base model modification is that 3D data of the extremal boundary (or the sillouette) and the center line (in the 3D sense) in the front view are obtained. If the other features, such as nose, eyes and mouth, are extracted and matched, the corresponding vertices in the model are modified; if they are not extracted and matched, the corresponding vertices in the model are treated the same way as those non-feature ones. The positions of non-feature vertices in the model are determined as a linear function of the positions of boundary vertices.

4. Knowledge-Based Facial Image Processing

Let us first describe assumptions on the stereo images. The two images are one front view and one side view of the face, with the angle between the two optical axes being 90 degrees. Long focal length is used to approximate orthographic projection. Both image planes are vertical and at the same height so that the projections of a space point onto the two images have the same vertical coordinate. This geometry guarantees that the three-dimensional coordinates of a space point (X,Y,Z) can be obtained by simply reading the horizontal and vertical coordinates of the corresponding points (xl,yl) and (yr,zr) in the two images (Figure 3) as

$$\begin{aligned} X &= -xr; \\ Y &= xl; \\ Z &= yl = yr. \end{aligned} \tag{1}$$

The more you know, the less you need to search [Winston, 84]. The more knowledge you have about the face, the less the algorithm complicates. Thus to facilitate processing we need to make full use of the knowledge about the face structure. In the following we make some assumptions about the face orientation and face structure.

(1) The head is vertical so that both the center line in the front view and the center line (an approximate one, defined as the line closest to the ear on the face side) in the side view are vertical.

(2) The face is oriented directly toward the front camera so that its image in the front view is symmetrical with respect to the center line (Figure 4.)

(3) A profile model of the extremal boundary in the front view and a profile model of the front half extremal boundary in the side view are prepared. They are compared with the edge images, and the differences are used to modify the model to approximate the edge image again (Figure 5.)

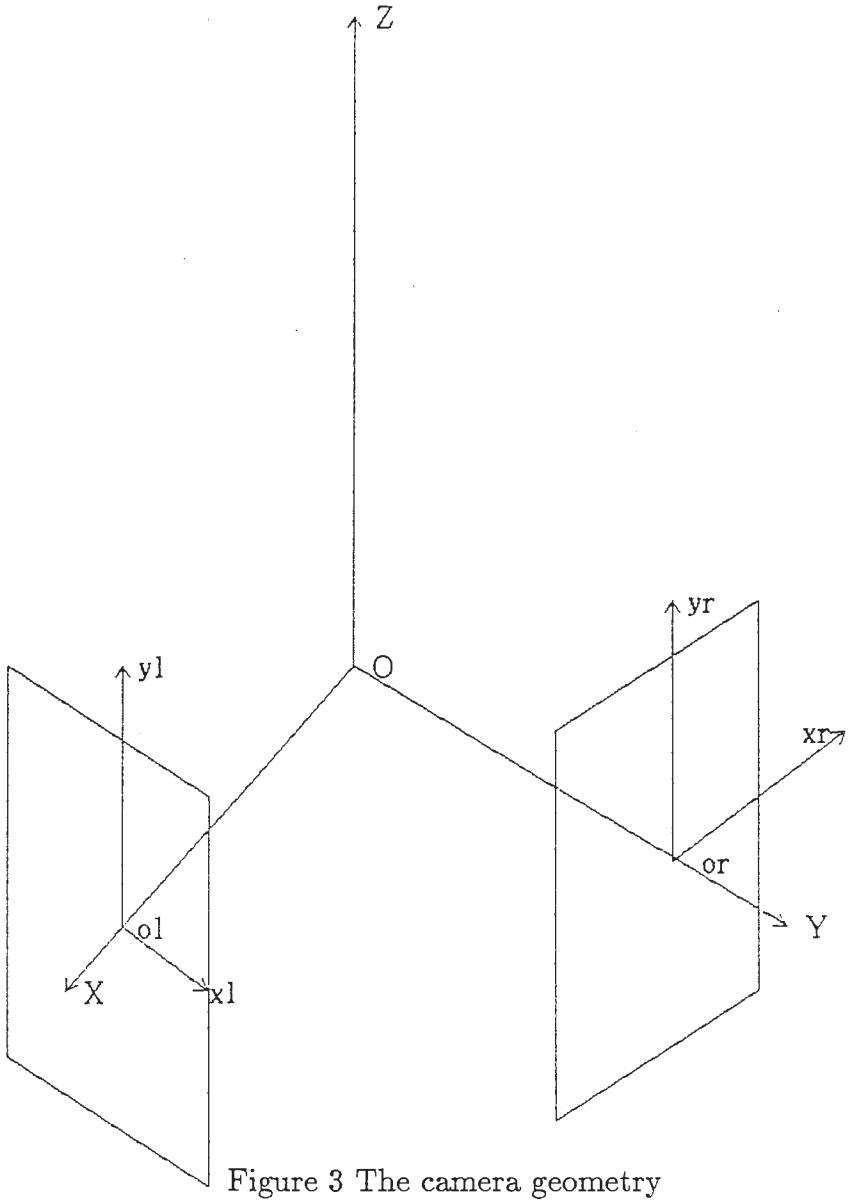

Figure 3 The camera geometry

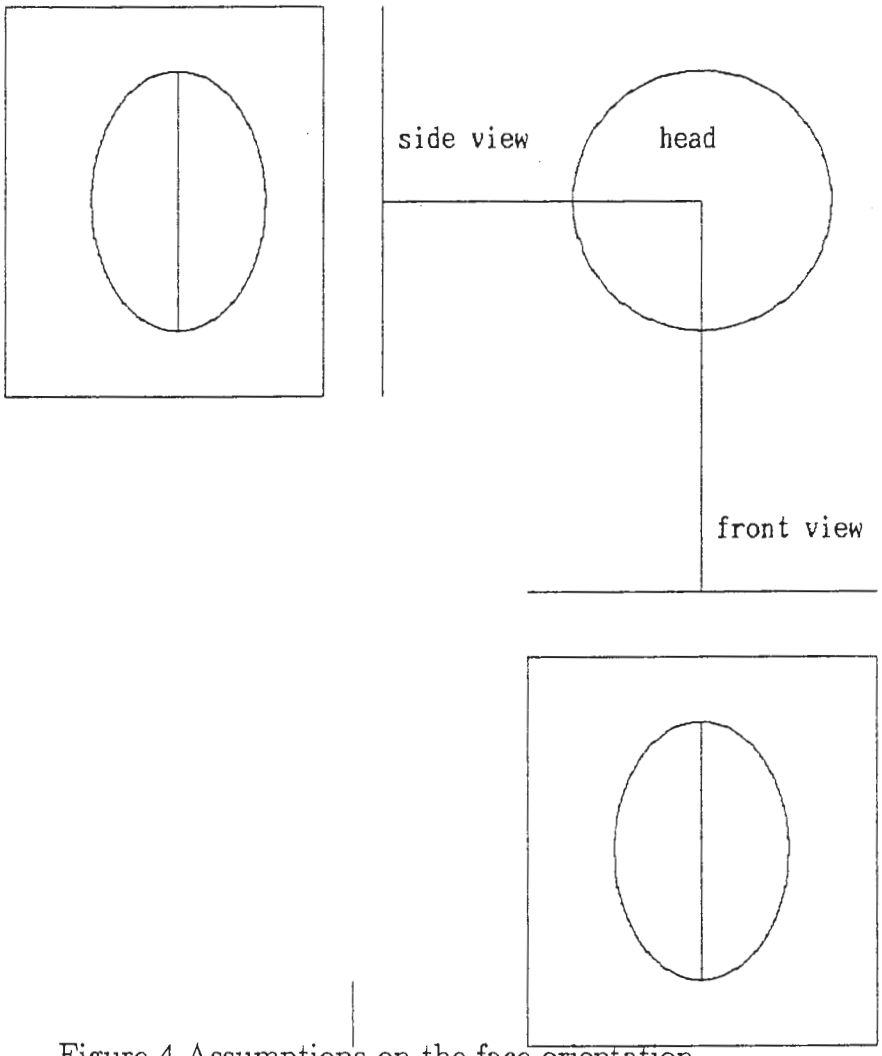

Figure 4 Assumptions on the face orientation

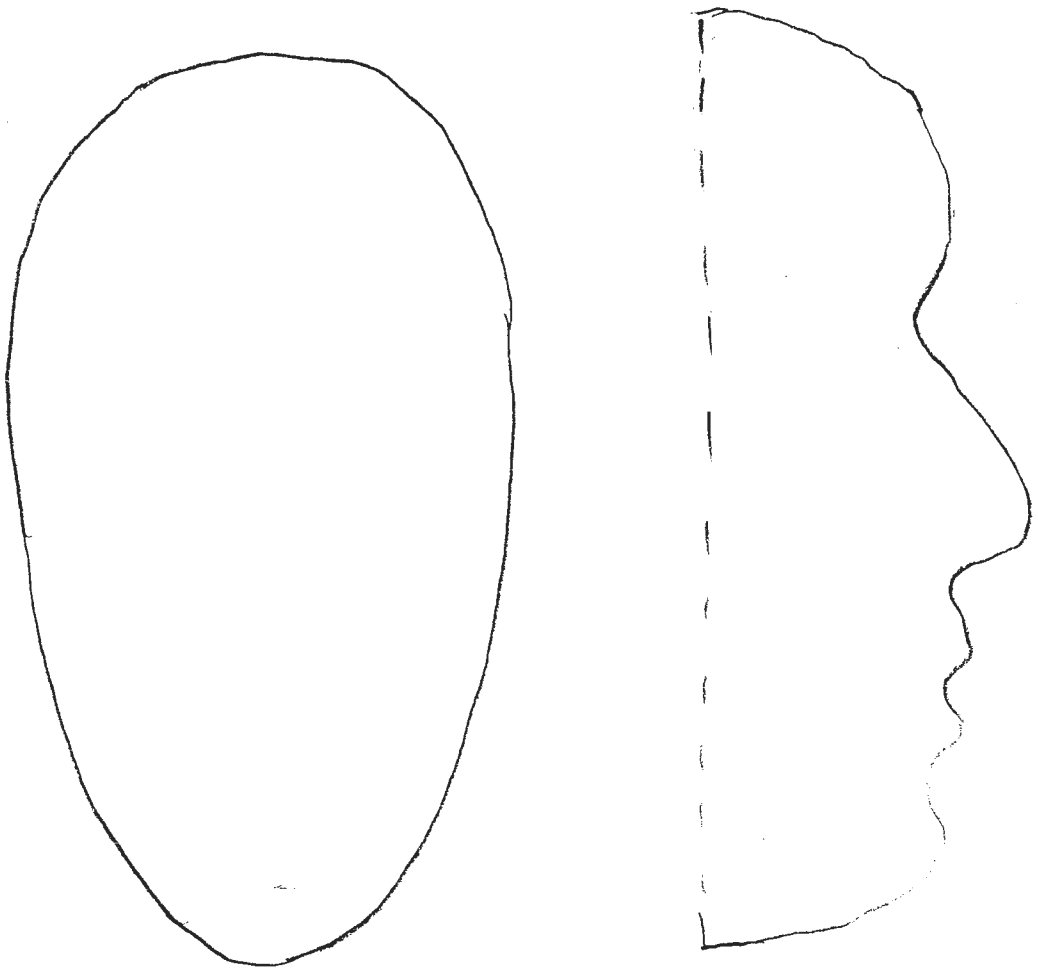

Figure 5 Profile models for the extremal boundaries

(4) A model of center positions of the features of eyes, nose and mouth in the front view is prepared to give an estimate to facilitate locating the features (Figure 6.)

The first step is to detect intensity changes with a differential operator, a Sobel or a Laplacian. A relatively low threshold is selected to ensure that even weak changes are picked up. The side effect is, of cource, that there will be more noise edges mingled with the true ones. We consider that this is preferable because that it is easier to delete undesired noise edges than to search for the desired edges that are lost.

As discussed in Section 3, the minimal requirement for the base model modification is that the extremal boundary (or sillouette) in the front view and the front half extremal boundary in the side view are extracted and matched. Two profile models are prepared and compared with the edge images. The profiles are generated by interpolating two sets of feature points with B-splines. The differences are used to modify the feature points which in turn generate new profiles. They are compared with the edge images again. After several iterations, the obtained profiles closely approximate the input extremal boundaries.

Once the profiles are extracted, the model of positions of eyes, nose and mouth is applied to the edge image to give a first-guess. Special filters are then used at the estimated locations to extract the features. Even using the model to restrict the search range, it is still not guaranteed that the features are extracted. Since the face is symmetrical with respect to the center line, if half features are extracted, we can mirrorcopy them as the other half.

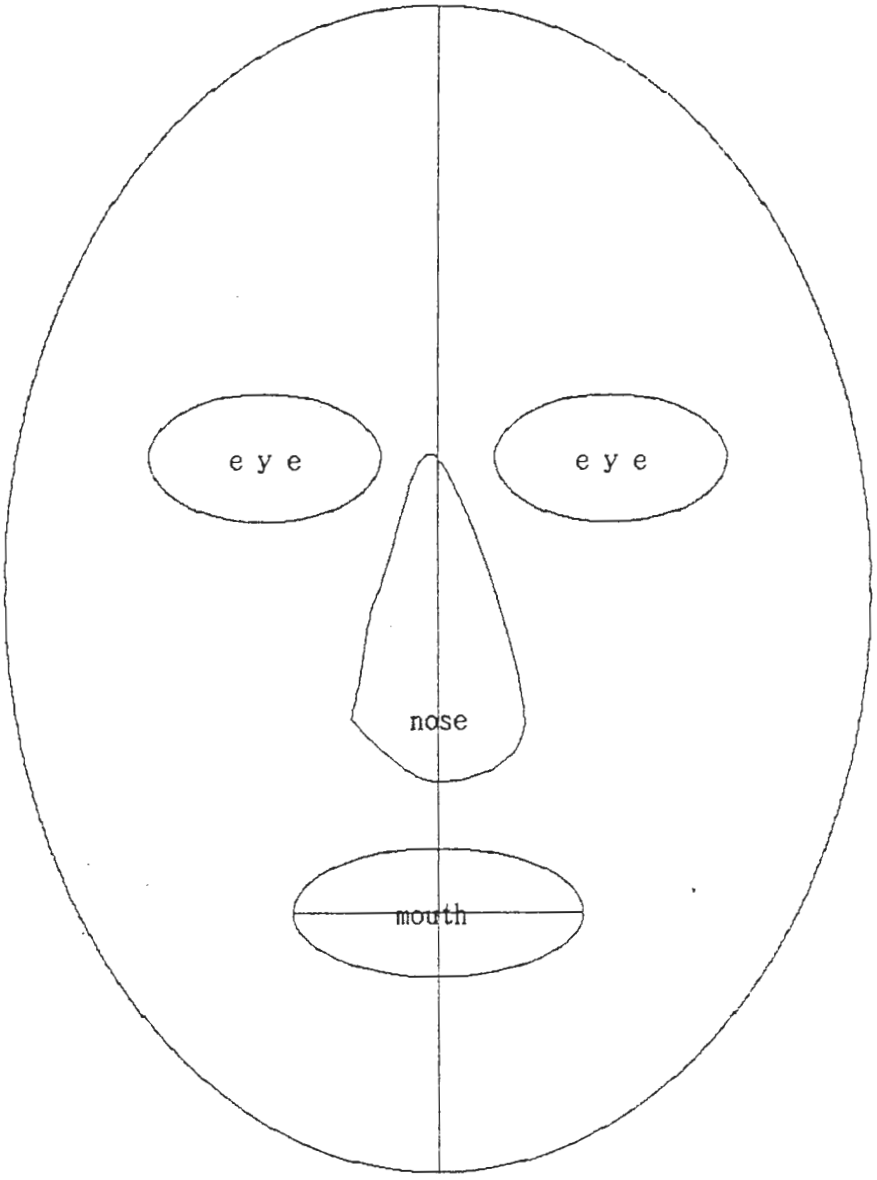

Figure 6 Position model for eyes, nose and mouth in the front view

5. Base Model and Its Adaptation to Acquired 3D Data

What we describe in the following is our considerations; the final result will inevitably be different to a certain extent from the current plan, due to some details overlooked here.

The base model is composed of connected triangles, which only cover the front half of the head surface. Since the 3D data are obtained from stereo matching, the disparity information is of a horizontal nature. It is thus desirable for us to accomodate the vertex distribution or vertex selection to this nature. One possibility is that the head are sliced horizontally at a number of characteristic heights, and the horizontal contours are further divided into segments. The resulting segmentation points are used as the vertices.

The vertices are divided into three groups: the boundary vertex group, the feature vertex group and the non-feature vertex group. The boundary vertex group includes vertices on the extremal boundary and the center line. They are determined directly from the 3D data whose acquisition is guaranteed in the stage of image processing and stereo matching. Both the feature vertices and the non-feature vertices are inside the boundary approximated by connecting the neighbor boundary vertices. The positions of the feature vertices will be modified corresponding to the 3D data, if any, obtained in the stereo matching. If 3D data are not available for them, then they are treated as non-feature vertices, whose postions are determined as a linear function of the boundary vertices on the same horizontal line. Suppose that there is a non-feature vertex **A** (x1',y1') in the base model, and the leftest and rightest boundary vertices **Bl** and **Br** on the same height have the coordinates (xl',yl') and (xr',yr'), respectively. If the real positions of **Bl** and **Br** are determined as (xl,yl) and (xr,yr), respectively, then the real position of **A** is determined as

$$x1 = xl + (xr - xl)\frac{x1' - xl'}{xr' - xl'}; \tag{2}$$

6. Facial Image Synthesis

The final step is to synthesize facial images virtually viewed from different angles. Now we a triangular polygon model and two images of the face at hand. Given an arbitrary virtual viewing angle, we first project orthographically the model onto the image plane associated with that viewing direction.

The first question that arises is which triangles are visible and which are not. Invisibility of a triangle can be caused by either that its orientation turns away from the viewing direction, or that other triangle(s) stand before it, wholly or partially. The first case can be identified by simply calculating the triangle's orientation. The second case needs more computation. The distance information is necessary. The more distant triangles give up priority to the closer ones.

Once the vertices are projected onto the image plane associated with the new viewing direction, the pixel intensity values are mapped by referring to the original images. Two questions arise: to which image and to which pixel in it intensity is referred. The answer to the first question is that the image in which the triangle has a larger area is referred to by that triangle. Since the original two viewing angles are 90 degrees apart, the answer can be paraphrased as that the image whose viewing direction has a smaller difference with the triangle's orientation is referred to, because the smaller the difference between the viewing direction and the triangle orientation, the larger the area of the triangle's image will be (proportional to the cosine of the orientation difference.) Once the reference image has been determined, the corrrespondences are built between the pixels inside the triangle in the synthesis image and the pixels in the reference image. Backprojecting a pixel in the synthesis image onto the reference image, the closest pixel is selected as the correspondence whose intensity is passed to the pixel in the synthesis image. There can be many-to-one correspondence relations (many pixels in the synthesis image to one in the reference image.)

It is not true that no problem exists in this kind of intensity inheritance, because

the intensity received by eyes varies with location. Viewed from a new angle, the intensity value of a pixel is definitely different from the corresponding points in both original images. But it is true that it does be a way, because one has to somehow inherit the information from the original two images. An alternative can be to calculate a weighted sum from the intensities of the corresponding pixels in the two original images.

6. Concluding Remarks

We have in Part One presented a detailed survey of the papers dealing with facial image processing, facial modeling and facial animation. In Part Two, we have proposed a stereo-based approach to facial modeling, which is an integral component of the ATR virtual space conferencing system project initiated in the AI department. In summary, a front and a side views of the face are taken, and boundaries and features are extracted by making full use of knowledge about the face structure. Matching the boundaries and features, we have their 3D position data. A base face model is prepared and modified according to the acquired 3D data. Finally images are synthesized assuming new virtual viewing angles.

The implementation of the proposed idea is currently under way. Inevitably, some problems have been overlooked in this technical report, and some others have intentionally not been included. For example, one of the problems that remains open is the modeling and processing of hair.

As the implementation of the system is over, we will write another technical report, which we hope will provide new progress and material.

References

Aizawa et.al. *Modeling a person's face and synthesis of facial expressions for use in a model-based synthesis image coding system*, **IECE, IE**87-2, 1987, pp.9-15

Aizawa, *Image motion analysis by synthesis based on a structure model*, **Proc. PCSJ**88, 1988, pp.75-76

Akimoto, et.al. *Expressive facial image generation by automatic shape modification*, **IECE, Graphiics and CAD**, 28-14, 1987, pp.119-125

Akimoto, et.al., *face model synthesis from front/side views and 3D base model*, **Proc. PCSJ**88, 1988, pp.69-70

Doyama, *Feature extraction for automatic identification of facial images*, **IEE-Yokoshu** 84-02-1, 1984, pp.1-6

Harashima, *Recent trends in analysis/synthesis coding systems for facial images*, **Television Society Journal**, Vol.42, No.6, 1988, pp.519-525

Kaneko, et.al., *Coding of face images based on 3D model of head and analysis of shape changes in input image sequence*, **IECE, IE**87-101, 1987, pp.79-86

Marr and Nishihara, *Representation*, **Artificial Intelligence**, Vol. 17, 1981

Muragami, et.al. *A study on image generation and transformation for human faces*, **IECE, IE**88-1, 1988, pp.1-8

Nemoto, et.al., *Synthesis system of 3d facial animation*, **Nikkei Computer Graphics Summer Issue**, 1986, pp.58-65

Noguchi et.al. *A simple modeling method for a facial shape*, **IECE Transaction D**, Vol.J71-D, No.11, 1988, pp.2350-2356

Numazaki, et.al., *Model-based identification of persons from facial images*, **IECE, PRU**87-122, 1987, pp.25-32

Parke, *Parameterized models for facial animation*, **IEEE Computer Graphics and Applications**, 1982 Nov., pp.61-68

Sakai, et. al., *Computer analysis of photographs of human faces*, **IECE Transaction**

D, Vol.56-D No.4, 1973, pp.226-233

Seki, et.al., *Feature points extraction for the human face picture*, **IECE, PRL**80-8, 1980, pp.1-8

Waters, *A muscle model for animating three-dimensional facial expression*, **ACM Computer Graphics**, Vol.21, No.4, 1987, pp.17-24

Winston, **Artificial Intelligence**, 1984, Addison-Wesley

Yang, et.al., *Model-based approximation of profile edge in human face recognition*, **IECE, PRU**86-125, 1986, pp.17-24