〔非公開〕

ＴＲ－Ｃ－００４７

最終報告書

徐　剛
XU　GANG

１９９０．３．２８

ＡＴＲ通信システム研究所

最 終 報 告 書

徐　剛

１９９０年3月

序

私は ATR に一年半ほど滞在させていただきました。最初の半年は研修研究員（その最初の数カ月は週一回程度の来所）として、後の一年は客員研究員として。途中、世の中も私自身も色々な事が起こったので、この一年半はだいぶ長い、そしてまた非常に短い、ような感じが致します。日本では、年号が換わった、中国では、天安門事件が起こった、ヨーロッパでは、ベルリンの壁が撤去され始めた。そして、私も北京大学に行く計画を改め、大阪大学に戻ることにしました。このように、毎日目を新聞から離せない中、 ATR での滞在期間は三回（身分変更一回含めて）も延長しました。とにかく、困難な時に居場所を提供して下さった ATR の皆様に深く感謝致します。

さて、研究の話に戻りますが、私は、臨場感通信、仮想空間通信会議システムの一部となっている顔のモデリングと認識に関する研究を担当しました。遠隔地に居る各参加者の三次元モデルを仮想の会議室に写し込み、そして仮想の参加者間位置関係に一致した画像を生成することによって、臨場感を作り出すというユーニクな発想です。顔の３次元モデルは再生画像の質に直接関係することから、重要視されています。モアレ法や光切断法などの方法では、正確なモデルが得られるが、特殊の設備が要るので、あまり実用ではありません。従って、カメラからの入力画像から顔の３次元モデルを作ることが要求されます。これが私の最初の仕事でした。私のもう一つの課題は、顔表情の認識です。表情の認識ができれば、認識の結果だけ伝送して、受信側で３次元モデルに表情の変化を加えることにより、表情のある顔画像を再生できるという”夢”です。この夢が何時実現するか分からないが、顔表情の認識自身が一つ重要な研究課題であってきました。人間にとって非常に簡単であっても、計算機にとって非常に困難な一好例です。 von Neumann 型計算機でなく、ニューラルなアポローチについて考察を行ないました。最後、私の本来の分野であるコンピュータビジョンについても研究を続けています。内容は線画の解釈ですが、通信の立場からとらえても有意義であると思われます。画像を線で表現し、伝送して、それを受け取った側で解釈し理解するというように、知的通信の枠組にも入る。

本報告は以上の三つの研究の内容をまとめたものです。それぞれの間は必ずしも論理的関

係が存在するわけではありません。第一部は、顔の３次元モデル化に関するもので、電子情報通信学会論文誌Ｅへの投稿（阿川、永嶋、岸野、小林と共著）からなっています。第二部は、顔表情の認識に関するもので、テクニカルレポートとして提出する予定の原稿（永嶋、岸野と共著）からなっています。第三部は、線画の解釈に関するもので、International Journal on Artificial Intelligenceへの投稿（田中と共著）からなっています。第二部の内容については、実験を行なっていないので、構想の域に留まっている。今後、どなたかのお役に立てば、幸いと存じます。

謝辞

私の受け入れにあたってご協力下さった、葉原会長、山下社長、小林前室長、岸野室長に深く感謝します。

研究の面で色々ご協力下さった永嶋氏、阿川氏、その他の３ｄグループの皆さんに深く感謝します。

色々と教えて下さったDaniel（李さん）に深く感謝します。

私をATRに紹介して下さった、そして普段色々とお世話になった先輩の田中弘美氏に深く感謝します。電車がもっと速く走るように期待します。

最後に、ATR通信システム研究所の皆さん、知能処理研究室の皆さん、企画課の皆さん、ブリッジの仲間に入れて下さった皆さんに深く感謝します。もっと上達したらと思っています。

目次

第一部

顔モデリング及び顔画像の生成

title:

Three-Dimensional Face Modeling for the ATR Virtual Space Teleonferencing System

authors:

Gang Xu, non-member, Hiroshi Agawa, non-member,

Yoshio Nagashima, member, Fumio Kishino, member, and

Yukio Kobayashi*, member

affiliation:

Artificial Intelligence Department

ATR Communication Systems Research Laboratories

Seika-cho, Soraku-gun, Kyoto 619-02, Japan

* Current address: NTT Human Interface Laboratories

1-2356 Take, Yokosuka-shi, Kanagawa 238-03, Japan

acknowledgement:

The authors would like to thank Dr. Kohei Habara, Chairman of the ATR Governing Board, and Mr. Koichi Yamashita, President of the ATR Communication Systems Research Laboratories for their advice and encouragement.

running head:

3D FACE MODELING

ABSTRACT

The goal of this research, as an integral component of the ATR virtual space teleconferencing system project, is to generate three-dimensional facial models from facial images and to synthesize images of the models virtually viewed from different angles. Since there is a great gap between the images and a 3D model, we argue that it is necessary to have a base face model to provide a framework. The base model is built by carefully selecting and measuring a set of points on the face whose corresponding points can be readily identified in the input images, and another set of points that can be determined from the first point set. The input images are a front view and a side view of the face. First the extremal boundaries are extracted or interpolated, and the face features such as eyes, nose and mouth are extracted. The extracted features are then matched between the two images, and their 3D positions calculated. Using these 3D data, the prepared base face model is modified to approximate the face. Finally, images of the modified 3D model are synthesized by assuming new virtual viewing angles. The originality and significance of this work lies in that the face model can be automatically generated.

1. INTRODUCTION

The focus of research activities in the AI department, ATR Communication Systems Research Laboratories, is the ATR virtual space teleconferencing system project [Kishino & Yamashita, 1989; Ishibashi *et al.*, 1988; Xu *et al.*, 1989b]. The general aim of the project is to achieve maximum 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 the models into a virtual conference room, and finally to send back stereo images synthesized in consistency with the assumed spatial relations among the participants in that conference room (Figure 1).

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

Communication is one of the basic needs of human beings. The most basic 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. What telephone can carry is verbal information. But sometimes people need not only verbal 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 sometimes brings mental distance, causing difficulties in friendly exchanges of opinion and experience. To overcome these problems in current visual communication systems, or, to achieve maximum virtual presence, is exactly what we are pursuing in this 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 [Harman & Hunt, 1977; Sakaguchi *et al.*, 1989]. 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. In the ATR virtual space teleconferencing system that we are trying to build, face images are required to be much more realistic than the other parts of the conference participants. For the images to be realistic, the 3D models must be close approximations of the faces.

Face modeling is indispensible not only to our virtual space teleconferencing system, but also to other image synthesis technologies. Facial animation [Parke, 1982; Waters,

1987] needs a 3D facial model in advance. Recently, the concept of "intelligent picture coding" (also called "knowledge-based picture coding") has been much advocated by [Harashima, 1988a], and efforts [Harashima, 1988a,b; Aizawa *et al.*, 1988] have been made along this direction. Within this framework, pictures are not transmitted directly, but are first analyzed and modeled, and then transmitted. Since the both sides share the same models or the same knowledge, what the receiver needs to do is to fuse the transmitted variational information with the structural information included in the model. This strategy significantly reduces the amount of information needed to be transmitted. In the specific case of face, given a face model, expressions can be represented as local variations of the model. Receiving the variations on the other side, expressions are then synthesized.

So far a number of face modeling methods have been developed. Most of them [Aizawa *et al.*, 1988; Akimoto & Suenaga, 1988] have been techniques that manually manipulating a prepared face model to fit the current face with the model being projected and superimposed with a given face image, usually a front view, sometimes both a front view and a side view. Most of the models are represented by connected triangular patches. The shape that such a model describes can be changed by three-dimensionally moving the vertices while keeping the topology invariant. Recently, Noguchi *et al.* (1988) proposed a more accurate method, which first pastes hundreds of point markers on the face, and then measures their 3D coordinates by matching the markers in different images. The face points are finally connected to form a triangular patch model.

In our approach we also use a triangular patch model. A base face model is first built in a fashion similar with that of Noguchi *et al.* (1988). By coordinating various image processing techniques, the 3D positions of face boundaries and face features are obtained, and the base model is consequently modified according to the obtained 3D data, to approximate the face shape.

In the following, Section 2 presents a general discussion on the strategy taken in the approach. Section 3 describes the construction of the base face model and the adaptation

of the base model to acquired 3D data. Section 4 gives the image processing techniques employed to obtain face boundaries and face features. Section 5 discusses how to synthesize facial images virtually viewed from other angles given the updated 3D model. Section 6 describes the implementation and results. Finally, Section 7 summarizes the approach and discusses problems and solutions.

2. 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. For a face model, there are two choices. One is the generalized cylinder representation, with cross sections being ellipses and the nose being approximated by small triangles attached to the ellipses (Figure 2). 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 (Figure 3). It has two layers of information, the topology and the vertex positions. To describe a face, the number of vertices is usually above 400. Comparing the two, one finds that the number of parameters to be specified in the second representation is greater than that in the first one. 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 patch model is much more desirable in our case, because a generalized cylinder with elliptic 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 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 to extract face features robustly. However, one has to admit that whatever methods 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 the use of a base model. That is, a base face model is first built from a typical face, and it is then modified to approximate a new face according to the 3D data acquired by matching the front and side views of that face.

3. THE BASE FACE MODEL AND ITS ADAPTATION TO ACQUIRED 3D DATA

The base model is composed of connected triangles, which 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 [Xu *et al.*, 1989a]. It is thus desirable for us to accomodate the vertex distribution or vertex selection to this nature. Our idea is to slice the head 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 neighboring vertices are then connected to form a triangular patch surface approximating the face.

The qualification for the characteristic heights is that they be steadily and easily recognized. It is evident that these heights are best illustrated along the profile in the side view of the face [Harmon & Hunt, 1977]. Along the profile, the following points are used as characteristic heights to slice the head in the top-down direction: eyebrow, nose bridge, nose tip, nose bottom, upper lip, mouth, lower lip, under-lip dimple and chin tip (see Figure 4). Another characteristic height that does not appear on the profile is the height of eyes, which can also be easily and steadily recognized in the front view. Now the number of chartacteristic heights is 10.

Besides these characteristic heights, the profile is further sliced into vertical intervals of nearly equal lengths. For the current model as shown in Figure 4, there are totally 24 heights, including the top of the head and the bottom of the face. The horizontal contours at these heights are further segmented into arcs of nearly equal lengths, with less segments near the top of the head and the bottom of the face. In the current experiment, we assume that the face is symmetrical with respect to the center line. Thus we need only to measure the shape of one of the two halves of the face; the other half can be mirror-copied. The vertices are marked by attaching the small circular paper pieces to them. And then the three-dimensional positions are measured of these markers by taking a front and a side views and matching the markers between the two images. The geometry is the same as that described in Section 4. The number of the resulting vertices is 234 for the left half face. Adding the other half and substituting the common center line, the total number of vertices is 444. They form a network of 840 triangles. The wireframe shown in Figure 3 is a front view of this model.

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. Both the feature vertices and the non-feature vertices are inside the boundary approximated by connecting the neighboring boundary vertices. The positions of the feature vertices will be modified corresponding to

the 3D data 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. This flexibility guarantees that the system would not be paralyzed by a single local failure in extracting a face feature, though this does not imply that the system frequently fails to extract them.

Suppose that the positions of two vertices $\mathbf{B}_l$ and $\mathbf{B}_r$ on some height are known as (x_l, y, z_l) and (x_r, y, z_r), respectively, and the corresponding vertices $\mathbf{B}'_l$ and $\mathbf{B}'_r$ in the base model have the coordinates (x'_l, y', z'_l) and (x'_r, y', z'_r), respectively. Then the position of vertex $\mathbf{A}$ (x, y, z) in between $\mathbf{B}_l$ and $\mathbf{B}_r$, whose corresponding vertex in the base model has the coordinates (x', y', z'), is determined as

$$\begin{aligned} x &= x_l + (x_r - x_l)\frac{x' - x'_l}{x'_r - x'_l}; \\ z &= z_l + (z_r - z_l)\frac{z' - z'_l}{z'_r - z'_l}. \end{aligned} \tag{1}$$

$\mathbf{B}_l$ and $\mathbf{B}_r$ can be either boundary vertices or feature vertices. If there are no feature vertices on the height, then they are boundary vertices.

4. FACIAL IMAGE PROCESSING AND 3D DATA ACQUISITION

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 so that the face's size is the same in the two images. 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 combining the horizontal and vertical coordinates

of the corresponding points (x_f, y_f) and (x_s, y_s) in the two images (Figure 5) as

$$X = x_f;$$
$$Y = y_f = y_s;$$
$$Z = x_s. \tag{2}$$

The eyes look horizontally toward the front camera so that the centerline in the front view is vertical and becomes an extremal boundary in the side view, which is usually called a *profile*. The background is set to be white so that the face area, the hair area and the background can be easily separated from each other (extraction of body area and face area from background under general conditions has been studied by another ATR group [Ishibashi *et al.*, 1988].)

Facial image processing has been previously studied [Doyama *et al.*, 1984; Sakai *et al.*, 1972; Seki *et al.*, 1980], but all the systems are more or less *ad hoc*. In the following we describe our image processing process, which consists of 4 steps.

(1) image segmentation and boundary extraction

As stated in last section, one prerequisite to the base model modification is that the boundaries are extracted and matched. In order to extract the extremal boundaries, which are boundaries between the face area and the background, we first take a histogram of each image. The two valleys between the areas are found to segment the image into regions, and then their boundaries are extracted. The boundary between the face and the background in the side view is the profile, which is projected as a straight line, the *centerline*, in the front view.

(2) feature extraction along the profile

Along the profile in the side view, we want to identify the eyebrow, nose bridge, nose tip, nose bottom, upper lip, mouth, lower lip, under-lip dimple and chin tip. The first characteristic of them is that they are all curvature extrema. The second characteristic is that all of them are local extrema of x-coordinate. Another discriminating criterion is the relative distances between the neighboring features. The nose tip is first identified as the right-most point along the profile.

(3) extraction of eyes, nose and mouth

Having obtained the vertical positions of the eyes, nose and mouth, we proceed to identify the positions in the front and side views of the eyes' left and right ends, the nose's left and right ends, and the mouth's left and right ends (see Figure 13(c).) Note that the inner ends of the eyes are not available directly but are estimated instead. The position data obtained in Step 2 are used to open windows at the areas so that the search can be separated and restricted to the right ranges. The images are first filtered with a Laplacian operator, and the filtered image is thresholded. Cutting it with the windows and applying thinning techniques to each window yields curve segments that approximate the boundaries of the face features. Finally, tailored algorithms are executed to locate the end points and intersection points of the curve segments, which are the feature points we look for.

(4) fitting of extremal boundaries

The output of Step 1 does not itself provide whole extremal boundaries that are necessary to modify the base face model. In the front view, both hair and neck hide or blur the extremal boundary. In the side view, hair covers the forehead above the eyebrow (see Section 6 for details.) Therefore, to recover the whole boundaries, we have to employ the base model in a two-dimensional manner. Endowing the extremal boundary of the base model with freedom of scaling in both the x- and y- directions, we find a best fit that has the least difference with the data of meaningful boundary segments obtained in Step 1. The profile of forehead in the side view is interpolated between the head top and the eyebrow while reserving the shape of forehead in the base model.

5. Facial Image Synthesis

The final step is to synthesize facial images virtually viewed from different angles. Now we have a triangular patch 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) stands before it, wholly or partially. The first case can be identified by simply calculating the triangle's orientation. The second case needs more computation because 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 are asked: to which image, and to which pixel in it, is the intensity 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, if it is visible in two images. The answer can be paraphrased as that the image whose viewing direction has a smaller difference with the triangle's normal is referred to, because the smaller the difference between the viewing direction and the triangle orientation, the larger the area of the triangle's projection will be in that image (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 (recall that the image is digital) 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 the eyes' location. Viewed from a new angle, the intensity value of a pixel is definitely different from the corresponding points in the 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. For example, the equation can be

$$i_s = \sqrt{i_{r1}^2 \cos^2 \alpha + i_{r2}^2 \sin^2 \alpha}, \tag{3}$$

where α is the angle between the new viewing direction and that of the first reference image. Note that all the above discussions are based on the assumption that the new viewing direction is constrained to lie on the XZ plane.

6. IMPLEMENTATION AND EXPERIMENTAL RESULTS

The proposed appproach has been implemented on a Sun3 workstation, a Vicom

image processor and an Iris graphics machine. We have conducted experiments on three persons, of which one is introduced in detail here. Figure 6 shows the front view and side view. The histogram of the side view is shown in Figure 7. Segmenting the image at the two valleys we get two binary images, shown in Figure 8, of which one shows the hair region (black area), and the other shows the head area. The face area is obtained by taking the difference of the two binary images; the boundary of the face area is shown in Figure 9(a). By similar processing, we have the boundary of the face area in the front view, which is shown in Figure 9(b). The profile in Figure 9(a) is segmented at eyebrow, nose bridge, nose tip, nose bottom, upper lip, mouth, lower lip, under-lip dimple and chin tip by finding out the extrema of curvature and x-coordinate, as shown in Figure 10(a). The other points along the profile that are not those feature points are interpolated while reserving the profile shape in the base model and are shown together with the features in Figure 10(b). The result of fitting the extremal boundary in the front view to the acquired data is shown in Figure 11, in which the solid line segments are the points that lie on the extremal boundary. Figure 12(a) shows the image obtained by filtering the front view with a 9×9 Laplacian operator and thresholding it at a suitable value. Figure 12(b) shows the window for the right eye, whose position is determined by referring to the result of profile segmentation. Applying the thinning techniques to the individual windows for each face features, we have the thinned curve segments in the front view and side view, as shown in Figure 13(a) and (b), respectively. Figure 13(c) and (d) show the extracted feature points superimposed with the original images, respectively. Combining the 2D data in the front and side views, we get the 3D position data for the boundary points and feature points, which are used to modify the base face model. The resulting model to approximately represent the face is shown in Figure 14(a) and (b), which are a front view and a side view of the wireframe, respectively. In the current stage, we have only mapped the front view to the wireframe model. Figure 15 (a) and (b) show one front view and one oblique view of the texture-mapped model, respectively.

7. CONCLUSION AND DISCUSSION

We have in the previous sections proposed a stereo-based approach to human face modeling and reported the implementation of the approach. The work described here is a part of the ATR virtual space teleconferencing system project, which aims at achieving maximal realism of visual presence in teleconferencing.

In summary of the approach, a front and a side views are taken of a human face, and boundaries and features are extracted by combining a number of image processing techniques. Matching the boundaries and features between the front view and the side view, we obtain their three-dimensional positions, which are used to modify a prepared base model to approximate the face. Finally images are synthesized by assuming new virtual viewing angles.

The implementation of the approach produced mixed results. While working well to two of the three tested faces, the system failed to extract the face boundary of the other one because his hair was long and blurs the boundary between the background and the face area under the ears. Therefore, it is one of our future tasks to generalize the assumptions and preconditions and to raise the system's robustness, because human faces vary quite largely from person to person. Another area which needs improvement is the quality of the base model. The current base model is too "handsome"; a more common face will be used to build the base model (it is also helpful to select one of the prepared base models that is the closest to the face.) The modified models are evaluated in terms of the degree of realism of the texture-mapped images. We consider that some sort of smoothing of the vertex positions in the vertical direction is helpful to produce softer synthetic facial images.

References

Aizawa, K., Yamada, Y., Harashima, H. and Saito, T. (1987) Modeling a person's face and synthesis of facial expressions for use in a model-based synthesis image coding system, **IECE Technical Report Vol. IE87-2**, pp. 9-15, in Japanese

Akimoto, T. and Suenaga, Y. (1988) Face model synthesis from front/side view and 3D base model, **Proc. PCSJ88**, pp. 69-70, in Japanese

Doyama, T., Okada, H., Nakamura, O. and Minami, T. (1984) Feature extraction for automatic identification of facial images, **IIEE-Proceedings Vol. 84-02-1**, pp. 1-6, in Japanese

Harashima, H. (1988a) Intelligent image coding and communication, **ITE Journal Vol. 42, No. 6**, pp. 519-525, in Japanese

Harashima, H. (1988b) Recent trends in analysis/ synthesis coding system for facial image, **ITE Technical Report Vol. 12, No. 9**, pp. 19-24, in Japanese

Harmon, L. and Hunt, W. (1977) Automatic recognition of human face profiles, **Computer Graphics and Image Processing, Vol. 6**, pp.135-156

Ishibashi, S., Akiyama, K. and Kobayashi, Y. (1988) A study of virtual space teleconference system and its human image processing, **IEICE Technical Report Vol. IE88-110,**, pp. 25-32, in Japanese

Kishino, F. and Yamashita, K. (1989) Communication with realistic sensations applied to teleconferencing system, **IEICE Technical Report Vol. IE89-35,**, pp. 1-6, in Japanese

Noguchi, N., Nakajima, M., Agui, T. and Fukuta, S. (1988) A simple Modeling method for a facial shape, **IEICE Trans. Vol. J71-D, No. 11**, pp. 2350-2356, in Japanese

Parke, F. (1982) Parameterized models for facial animation, **IEEE Trans. on Computer Graphics and Applications**, November, pp.61-68

Sakaguchi, T., Nakamura, O. and Minami, T. (1989) Personal identification through facial images using isodensity lines, **Proc. of SPIE'89 Conference on Visual Communication and Image Processing, Vol. 2**, pp. 634-645

Sakai, T., Nagao, M. and Kanade, T. (1972) Computer analysis and classification of photographs of human faces, **Proceedings of the first USA-Japan Computer Conference**, pp.55-62

Seki, Y. and Hashimoto, S. (1980) Feature points extraction for the human face picture, **IECE Technical Report Vol. PRL80-8**, pp. 1-8, in Japanese

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

Xu, G., Kondo, H. and Tsuji, S. (1989a) A region-based stereo algorithm, **Proc. 11th Interntional Joint Conference on Artificial Intelligence**, pp. 1661-1666

Xu, G., Agawa, H., Nagashima, Y. and Kobayashi, Y. (1989b) A stereo-based approach to face modeling for the ATR virtual space conferencing system, **Proc. of SPIE'89 Conference on Visual Communication and Image Processing, Vol. 1**, pp. 365-379

Biography

Gang Xu is a visiting researcher at ATR Communication Systems Research Laboratories. He received the B.E. degree from Nanjing Institute of Technology (now Southeast University) in 1982. He began his study at Osaka University in 1983, and received the M.Sc. and Ph.D. degrees in 1986 and 1989 (March), respectively. In recent years, he has published a number of papers in the field of computer vision and artificial intelligence.

Hiroshi Agawa received the B.E. and M.E. degrees in Electrical Engineering from Osaka University in 1984 and 1986, respectively. In 1986 he joined Kinki Nippon Railway Co., Ltd.. In early 1989, he joined ATR Communication Systems Research Laboratories, and is currently doing research on facial image processing for the ATR virtual space conferencing system.

Yoshio Nagashima was born in Chiba, Japan, in 1956. He received the B.E. degree from Ibaraki University in 1979. From 1979 to 1988, he joined NTT Electrical Communication Laboratories, where he was engaged in research and development of video communication terminals. In early 1989, he joined ATR Communication Systems Research Laboratories, and is currently working on facial image processing.

Fumio Kishino is the head of Artificial Intelligence Department, ATR Communication Systems Research Laboratories. He received the B.E. and M.E. degrees from Nagoya Institute of Technology, Nagoya, Japan, in 1969 and 1971, respectively. In 1971, he joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation, where he has been involved in work on research and development of image processing

and visual communications systems. In mid-1989, he joined ATR Communication Systems Research Laboratories, and became the head of Artificial Intelligence Department. He is a member of the Institute of Electronics, Information and Communication Engineers of Japan and the Institute of Television Engineers of Japan.

Yukio Kobayashi is the Executive Manager of Visual Perception Laboratory, NTT Human Interface Laboratories. He received the B.E., M.E. and D.E. degrees all from Touhoku University. From 1970 to 1986, he joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation, where he was active in work on evaluation of television picture coding techniques and standardization of videotex systems. In early 1986, he came to Osaka to establish ATR Communication Systems Research Laboratories with his colleagues, launched the ATR Virtual Space Conferencing System project, and served as the head of Artificial Intelligence Department till May, 1989. He is a member of the Institute of Electronics, Information and Communication Engineers of Japan and the Institute of Artificial Intelligence of Japan.

Figure 1 Virtual space teleconferencing

Figure 2 A generalized cylinder face model

Figure 3 A triangular polygon wireframe face model

Figure 4 The front and side views of the base face model

Figure 5 The camera geometry

Figure 6 The input images: a front view and a side view

Figure 7 The histogram of the side view

Figure 8 The segemented images

Figure 9 The face boundaries

Figure 10 Features and vertices along the profile

Figure 11 Boundary fitting for the front view

Figure 12 Laplacian-filtered and thresholded image and the window for the right eye

Figure 13 Feature extraction in the front and side views

Figure 14 A front view and a side view of the wireframe face model

Figure 15 A front view and an oblique view of the texture-mapped model

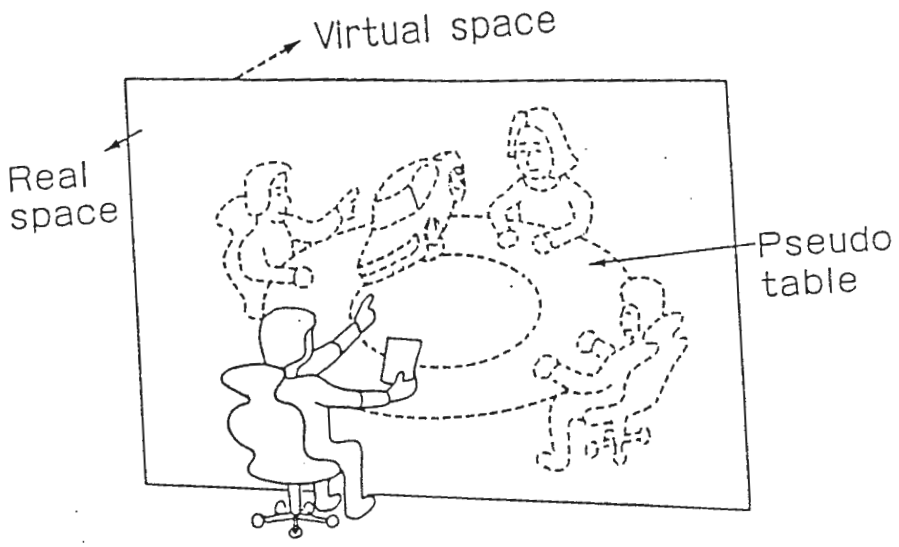


Fig. 1

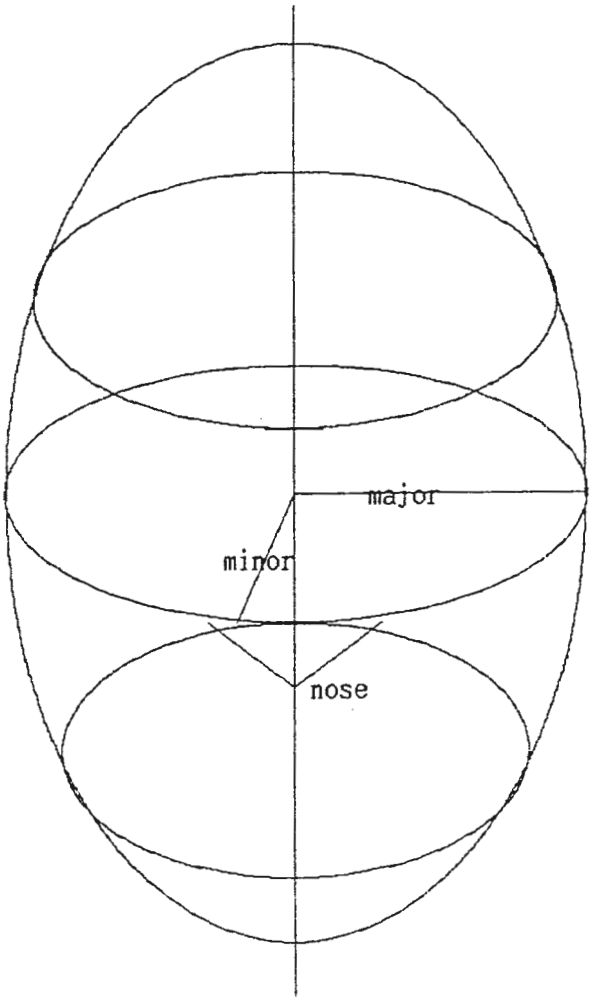


Fig. 2

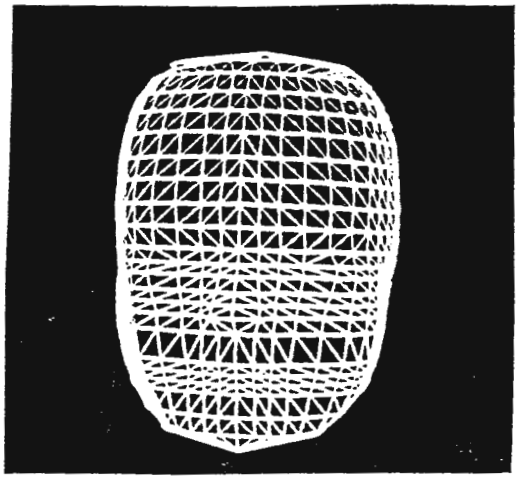


Fig. 3

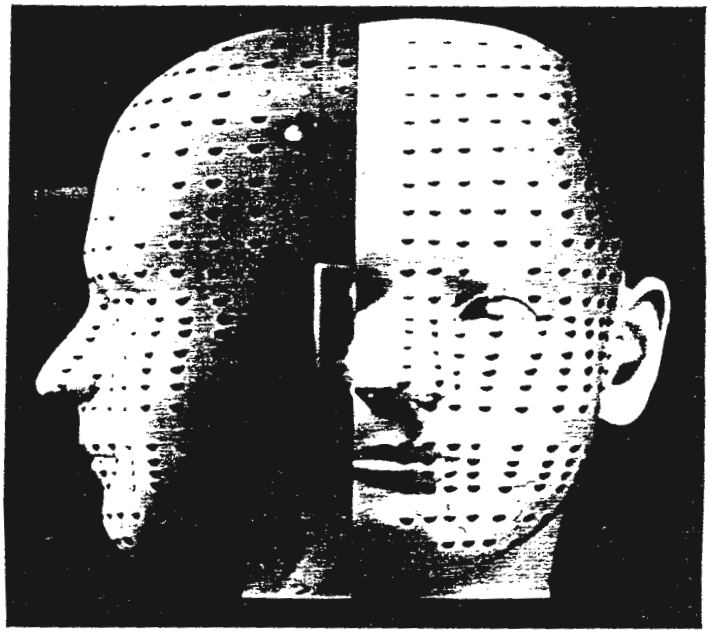


Fig. 4

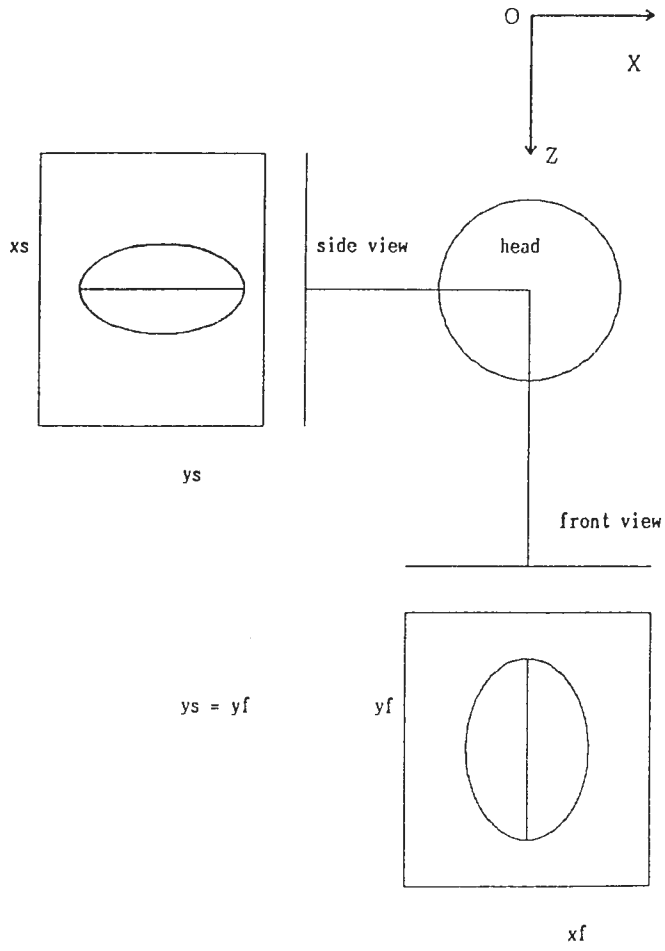


Fig. 5

(a)

(b)

Fig. 6

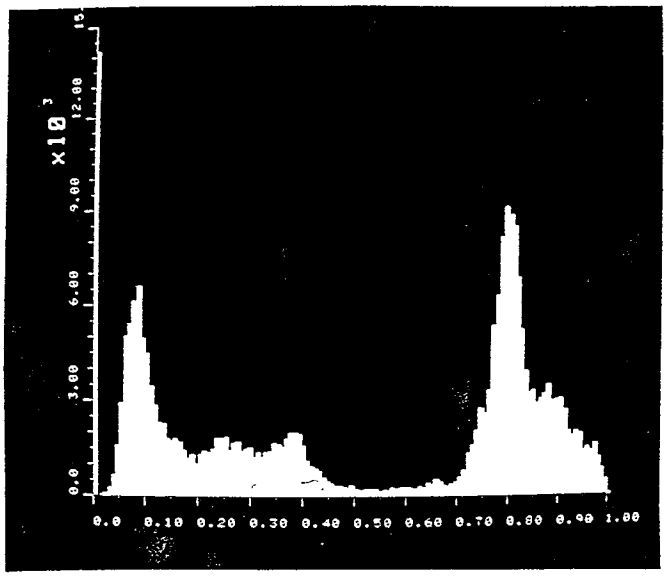
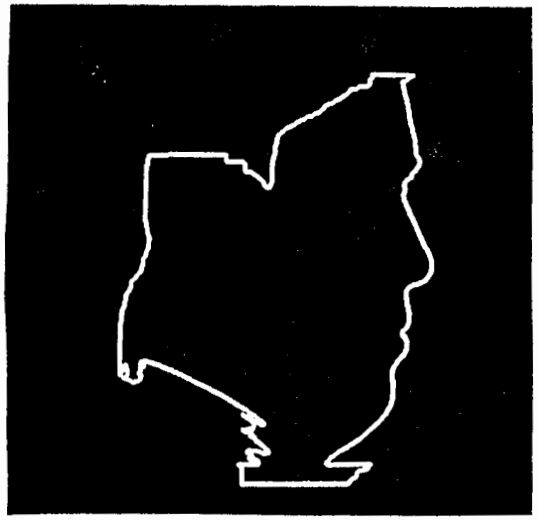

x10 3
15.
12.00
9.00
6.00
3.00
0.0
0.0 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00

Fig. 7

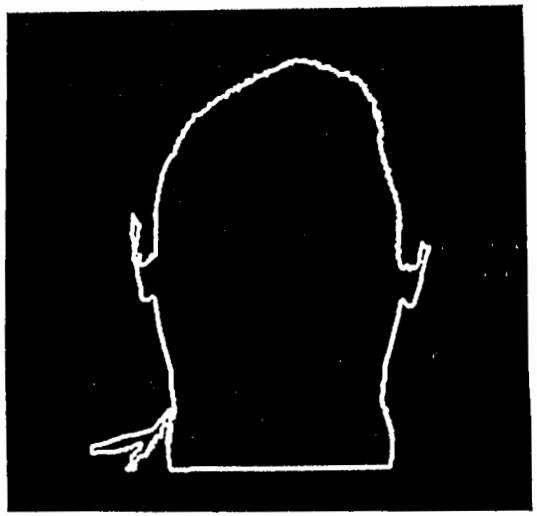
(a)

(b)

Fig. 8

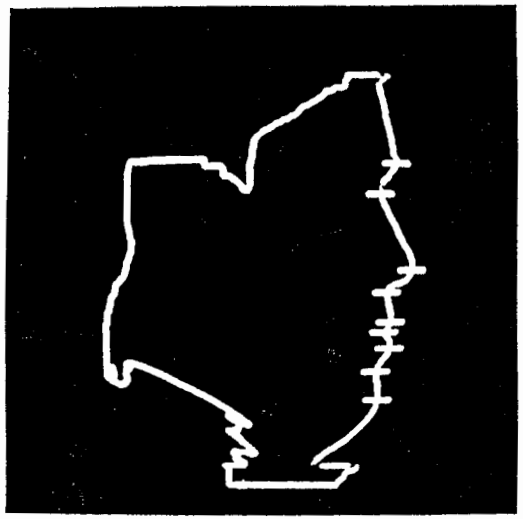


(a)

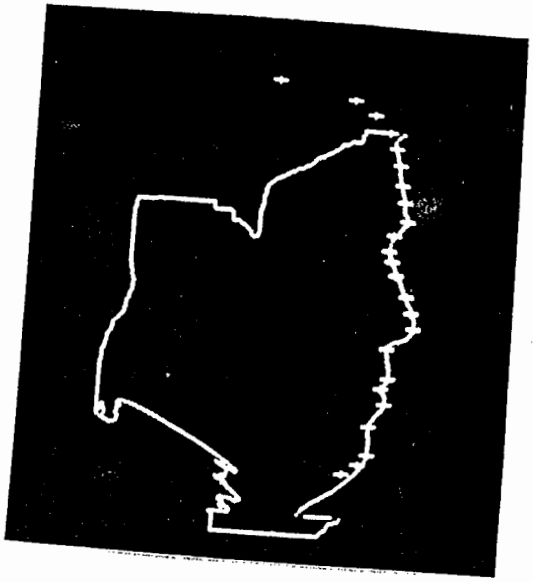


(b)

Fig. 9



(a)



(b)

Fig. 10

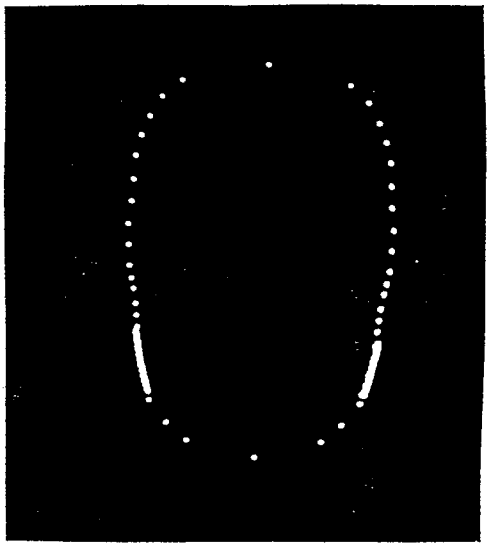
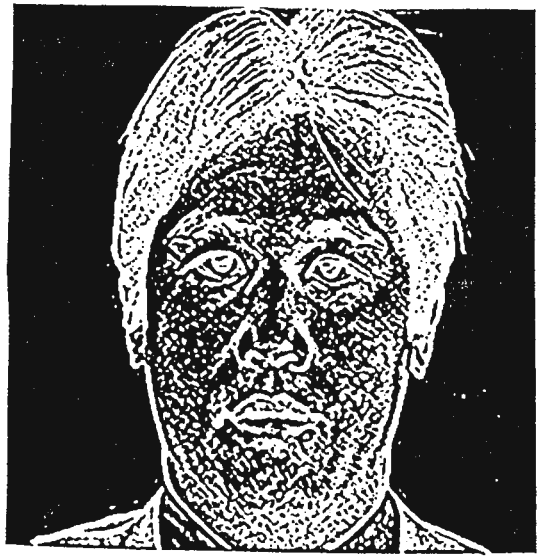
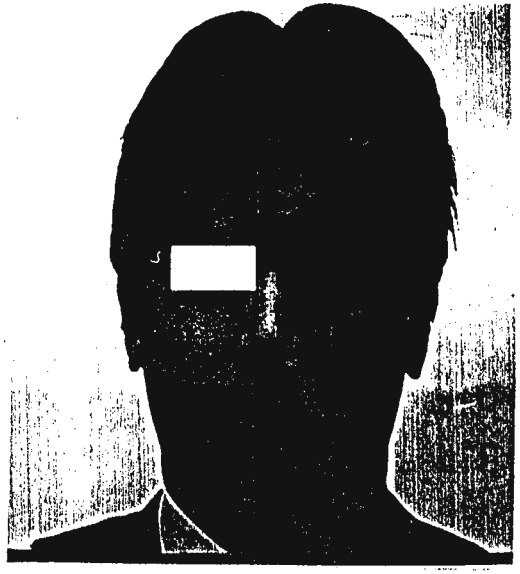


Fig. 11



(a)



(b)

Fig. 12

(a)

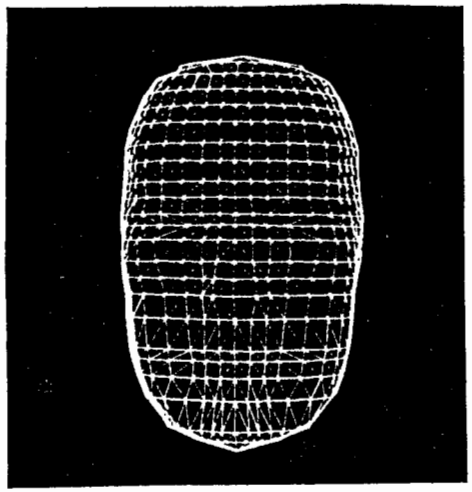
(b)

Fig. 13

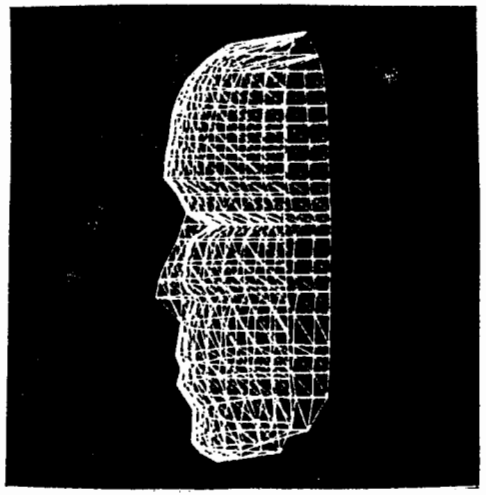
(c)

(d)

Fig. 13

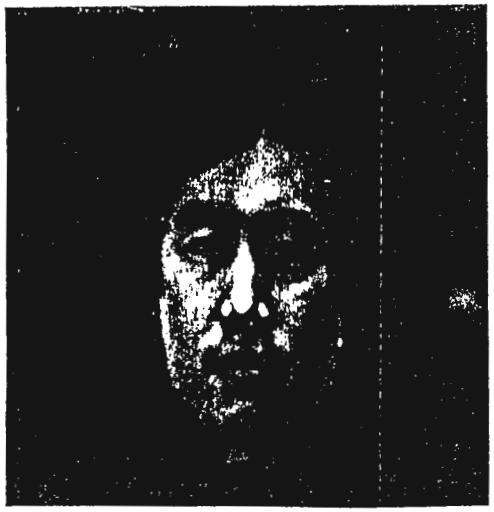


(a)



(b)

Fig. 14



(a)

(b)

Fig, 15

第二部

ニューラルネットによる表情認識に関する考察

ニューラルネットによる表情認識に関する考察

徐剛、永嶋美雄、岸野文郎

Abstract

This technical report presents the results of our investigation into the feasibility of a neural approach to facial expression recognition. Chapter 1 gives a general introduction, discussing the differences between conventional statistical and syntactical pattern recognition and neural pattern recognition. Chapter 2 gives a brief introduction of neural networks and pattern learning. Chapter 3 reviews previous work on neural face recognition systems. Chapter 4 proposes two models of facial expression recognition. Chapter 5 discusses matters related to implementation of the models.

第一章 まえがき

パターン認識は我々がたえず行なう行為である。本を読む時、文字の認識をしなければならない。人と会うとき、その人の顔を識別しなければならない。また、その人の表情をも認識して会話していく必要がある。

コンピュータにパターン認識の機能を持たせる試みが多くなされている。６２年ごろのパーセプトロン [1] 以外は、最近まで研究の主流は von Neumann 型計算機でパターンを定義し、入力をそのパターンの記述と照合することにより認識を行なうものであった。パターンの定義法は二種類ある。一つは統計的な定義法 [2] で、パターンをあるｎ次元の特徴値ベクトルで表現する。照合も入力とそのベクトルとの距離で行なわれる。この種のパターン認識は統計的パターン認識という。もう一つは構造的な定義法で、パターンを基本要素と基本要素の構造として表現する。照合は入力とパターンとの基本要素の構造間で行なわれる。この種のパターン認識は構文的パターン認識 [3] という。この二種類のパターン認識は六十年代より精力的に研究され、リモートセンシング、文字認識、指紋照合、医学などに広く応用されている。構造の記述

ができるので、表現力としては統計的定義法よりも構文的定義法の方が勝る。しかし、不安定な構造に対しては、構文的定義法も無能になり、むしろ統計的定義法の方がいい時が多い。いずれにしても、結論的に言って、 von Neumann 型計算機でパターン認識をする場合、パターンを予めある定義法に従って記述しておかなければならないが、詳細なおかつ包容的な”書き換え”が不可能なため、理想的なパターン認識が得られないことが多い。

ここへ来て、二三年前に、再びニューラルネット [4] のアポローチが見直され、丁度人工知能の行詰まりと重なっていることもあり、現在、熱い視線を浴びている。ニューラルネットによるパターン認識と統計的及び構文的パターン認識との違いは、前者はパターンを別の”言語”で書き換える必要がなく、パターンそのものを学習させ、覚えさせることにある。同じパターンのものを再び見せられる時、ネットは yes と出力する。従って、ニューラルパターン認識は特に顔とか顔の表情とかいう様な書き換えが難しいものに適しているといえる。

第二章 ニューラルネットと学習

確かに、ニューラルネットは発想が人間の神経組織から得ているが、両者は同じ構造を持っているわけではない。ネットは多くのセルよりできていて、各々のセルが近傍のセルと繋がっている。各セルが他のセルから入力を受けて、重みづけ和を計算し、それを一定の規則で変換して、次のセルに出力する。セル i のモデルを図 1 に示す。

$$o_i = f(\sum_j w_{ij} o_j)$$

ここで、o_i, w_{ij}, o_j はそれぞれセル i の出力、セル j との接続の強さとセル i へのセル j の出力である。$f()$ は非線形で、sigmoid 関数を使うことが多い。

現在、セルが層に分けられるネットがほとんどになっている。その中も、三層のネットが最も多い。最初の層は入力層で、真中の層は中間層といい、最後の層は出力層である（図 2 ）。セルは層と層の間つながりを持つが、同じ層の中では接続を持たない。このように、各セルを一個のプロセッサと考えて、ネット全体を並列分散計算機と考えてもよい。

パターン認識をさせるためには、まずパターンを学習させ、覚えさせなければならない。あるパターンに属する例をたくさん、繰り返して見せ、そして望ましい出力を得るようにセル

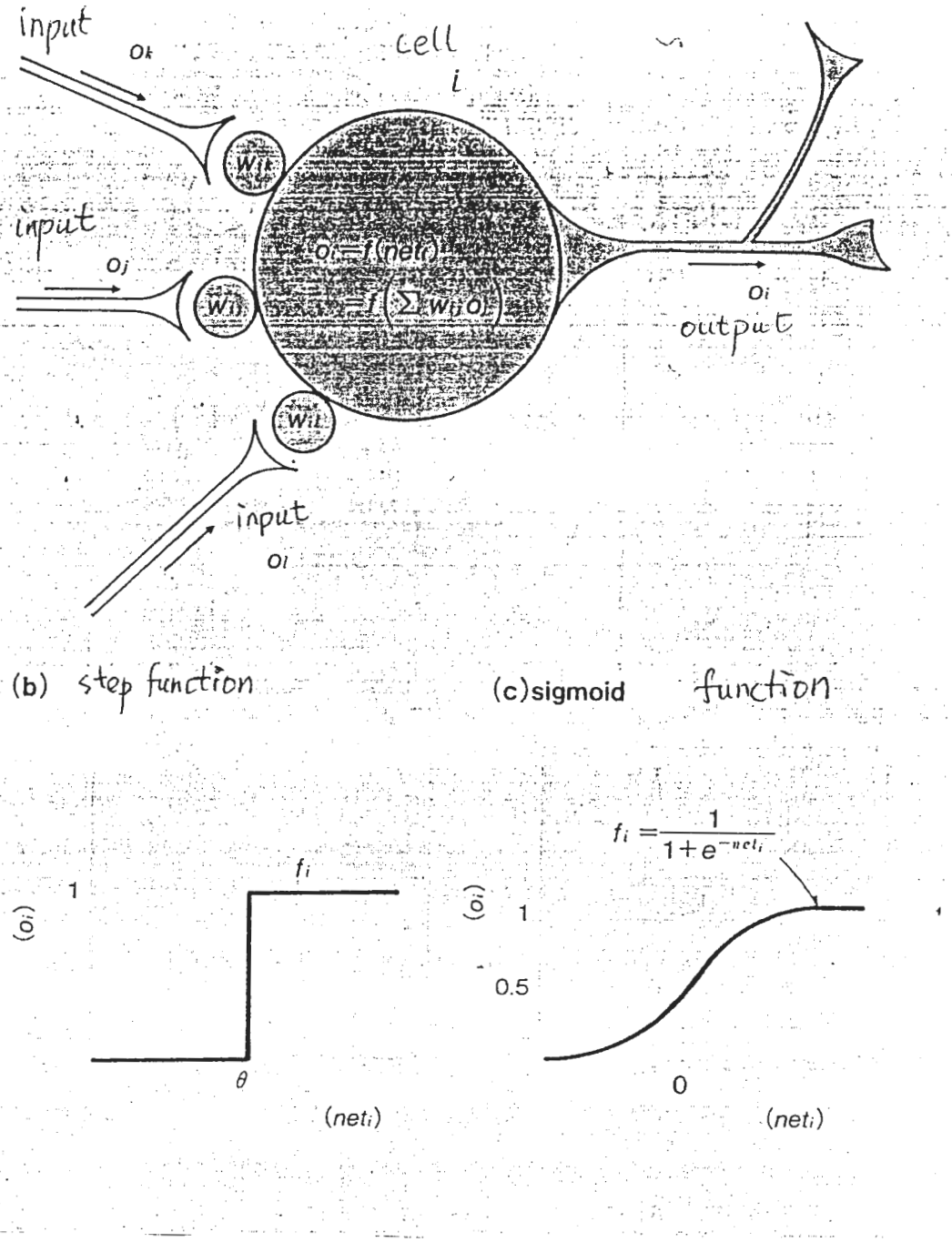


Figure 1 What is a cell?

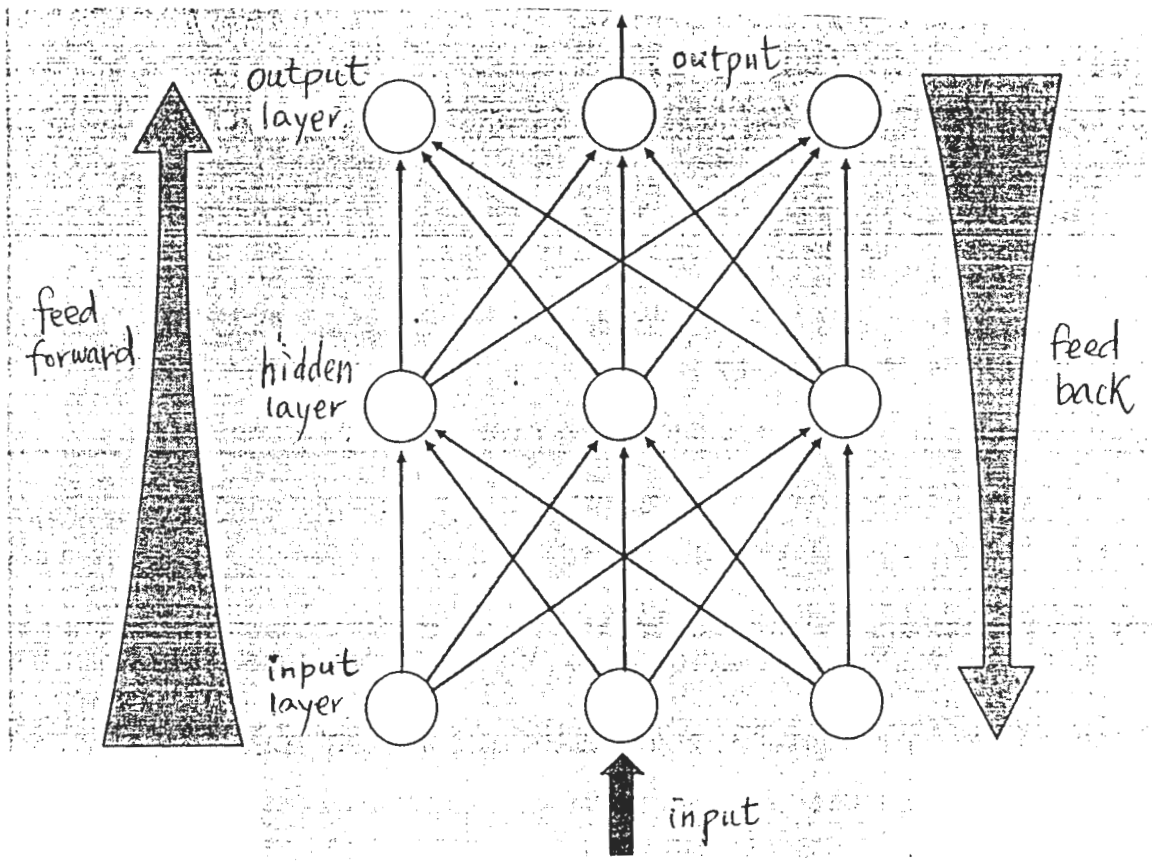


Figure 2 A three-layer network

とセルとの間の接続の強さを変えることにより、学習が行なわれる。学習法もいろいろあるが、よく使われるのはバックプロパゲーション法である。バックプロパゲーション法は1986年 Rumelhart ら [5] が提案したもので、簡単で汎用なため、現在一般的に使われている学習法となっている。入力層の各セルに入力をまず与える。そして出力層でその出力と望ましい値との差をネットに逆にフィードバックし、その差が無くなるようにセルとセルとの接続の強さを調整する。接続の強さの調整は一度で完成するのでなく、リカーシブに行なわれる。目標値を人間が与えることから、教師あり学習法ともいう。具体的には、次の三つの式によって、w_{ij} が変えられる。

$$\Delta w_{ij} = \eta \delta_i o_j$$

この式は w_{ij} の変化量が誤差 δ_i に比例し、そして w_{ij} を通じての入力 o_j に比例することを意味する。δ_i は出力層か中間層かで違う。出力層の場合は、

$$\delta_i = (t_i - o_i) f'(net_i),$$

t_i はセル i の目標値である。中間層の場合は、

$$\delta_i = f'(net_i) \sum_k \delta_k w_{ki}.$$

この式からも分かるように、誤差の出力層から入力層への後向き伝播は実は入力層から出力層への信号の前向き伝播と同じことをしている。w は一方通行でなく、両方通行である。前向き伝播と後向き伝播を一回のループとして、誤差がゼロに収束するまで繰り返す。学習例も一つだけでは十分な典型性を持たないので、できるだけ多くの代表性のある例を学習させることが重要である。この様に学習済みのネットは同じパターンのものを見せられる時、望ましい出力値に近い値を出力する。

第三章 ニューラル顔認識に関する従来の研究

報告されているニューラルパターン認識の研究の中、音声認識、数字認識、文字認識が多いが、顔認識の例はあまりない。原因は一つは、顔画像のサイズが大きいため、装置、学習時

間等の面で実際に試作することは大変である。二つ目は、応用が限られているため、まだ手をつける余裕はなかったかもしれない、と思われる。ここで入手できた例を二つほど紹介する。

まず、情報処理学会第36回全国大会で、成渓大学工学部の緑川氏[6]により、バックプロパゲーションによる顔画像認識の一考察との発表が行なわれた。顔画像はまず32×32のサイズに変換され、各ピクセルがそれぞれ1024の入力層のセルに接続される。中間層は4つか16個のセルから、出力層は4人の顔に対応する4つのセルからできている。学習は教師信号(1または0)と出力信号との差が0.1以下になるまで繰り返し行なわれる。四人の顔に対して学習が行なわれた。学習回数は4回、160回、2000回以上と初期重みに大きく依存する。これに対応して、学習時間も数分から数時間かかる。学習済みのネットに対し、ノイズ、ぼけ、ずれ、欠けを加えたパターンを提示した時の反応をも調べた。ノイズ、ぼけに強く、ずれに弱く、欠けは量や場所により認識の程度が異なることが分かった。

二つ目の例としては、1990年1月9日の読売新聞の報道[7]によると、日本電気は3500人の顔を一秒で認識できるシステムを開発した。新聞社に提出した紹介では、システムの特長として、四つ挙げている。一、高速である。二、メガねを掛けたり、外したりした顔でも正確に照合できる。三、並列プロセッサμPD7281を用いたため、顔学習、顔の位置合わせや入力画像の濃淡補正、拡大縮小などの前処理、照合の高速化が図れる。四、μPD7281を64個並列にして、320MIPSという高速処理を実現している。1人対100人の照合は0.12秒、1人対1000人の照合は0.6秒である。顔画像はまず32×32にサンプリングされる。8人の場合は入力層は1024個のセル、中間層は16個のセル、出力層は8個のセルからなっている。190回の学習の後、認識率は90%に達する。

二つの例を比較すると、速度と容量以外はあまり差はない。もう一つはいずれも、位置合わせ、拡大縮小、濃淡補正などの前処理が必要なことは残念に思われる。これに関して、福島等[8]のネオコグニトロンの方がスケール、位置の変化にも対応できることは、確かにネットを遥かに複雑にしているが、やはり助かることが多い。

この原稿を書いているところ、90年電子情報通信学会春期全国大会のある報告[9]を読んだ。画像パターンそのものを学習するのでなく、五つの特徴量を抽出し、その特徴量をニューラルネットに学習させ、認識を行なうというものであった。顔認識に適している特徴量の定義

ができれば、何もニューラルネットでする必要もないので、意味はない。

第四章 幾つかの提案

ニューラルネットによる表情認識に関する文献は見つからなかったが、次の問題にまず対応しなければならない。顔認識が表情の変化に左右されてはいけないと同様に、表情認識も顔が異なっても、笑いは笑いと判断できなければならない。言い替えれば、すべての人のすべての表情を記録しておけば、誰の何の表情かは認識できるであろうが、すべてのパターンを学習させる必要があるため、実用的でないし、人間の仕方に根拠を求めることもできない。そういう意味で、表情パターンは顔の個人パターンと違って、より抽象的であると言える。従って、その認識も顔の認識と異なる仕方で行なわなければならない。同じ問題は異なる話者の発音から単語を抽出、認識することにもある [10]。

この性質より、次のいくつかの方法が考えられる。

一）、顔の特徴点にマークを貼り、異なる表情の時のマークの動きベクトルを入力として、ネットに学習させる。入力層のセル数もベクトルの次元数と同じでよい。この場合には、スケール合わせが要るのが無論のことで、特徴点を選んで、マークを貼らなければならないため、前章の最後の例と同じ矛盾に陥ることになる。

二）、顔認識ネットに表情認識のネットを付け加える。顔認識によって平常時の顔パターンが得られる。表情顔との差をパターンとして認識することによって、表情を認識する。しかし、この場合は、依然としてすべての顔を登録し、学習させなければならない。

三）、学習したものをすべて覚えていて、一般化できると期待して、ただ学習例をたくさん提示する。これは賢い方法とはとても言えないが、ある程度の成功が期待できると思われる。

四）、顔の認識が輪郭や髪の毛など全体像が大きいウエートを持つのに対して、表情の認識は口や目の部分が決定的である。顔を口と目（二つの目を含む一つのパターン）の特徴的な部分パターンに分けて（図３）、各部分パターンについてニューラル認識を行ない、その総合的な結果として表情を認識する。この場合、スケール合わせが最初にできているとしても、各部分の位置決めが必要となる。なぜなら、他の例からも分かるように、ニューラルネットによる認識は位置ずれに弱いから。三）と比較して、より少ない学習例で済むことになるが、各人

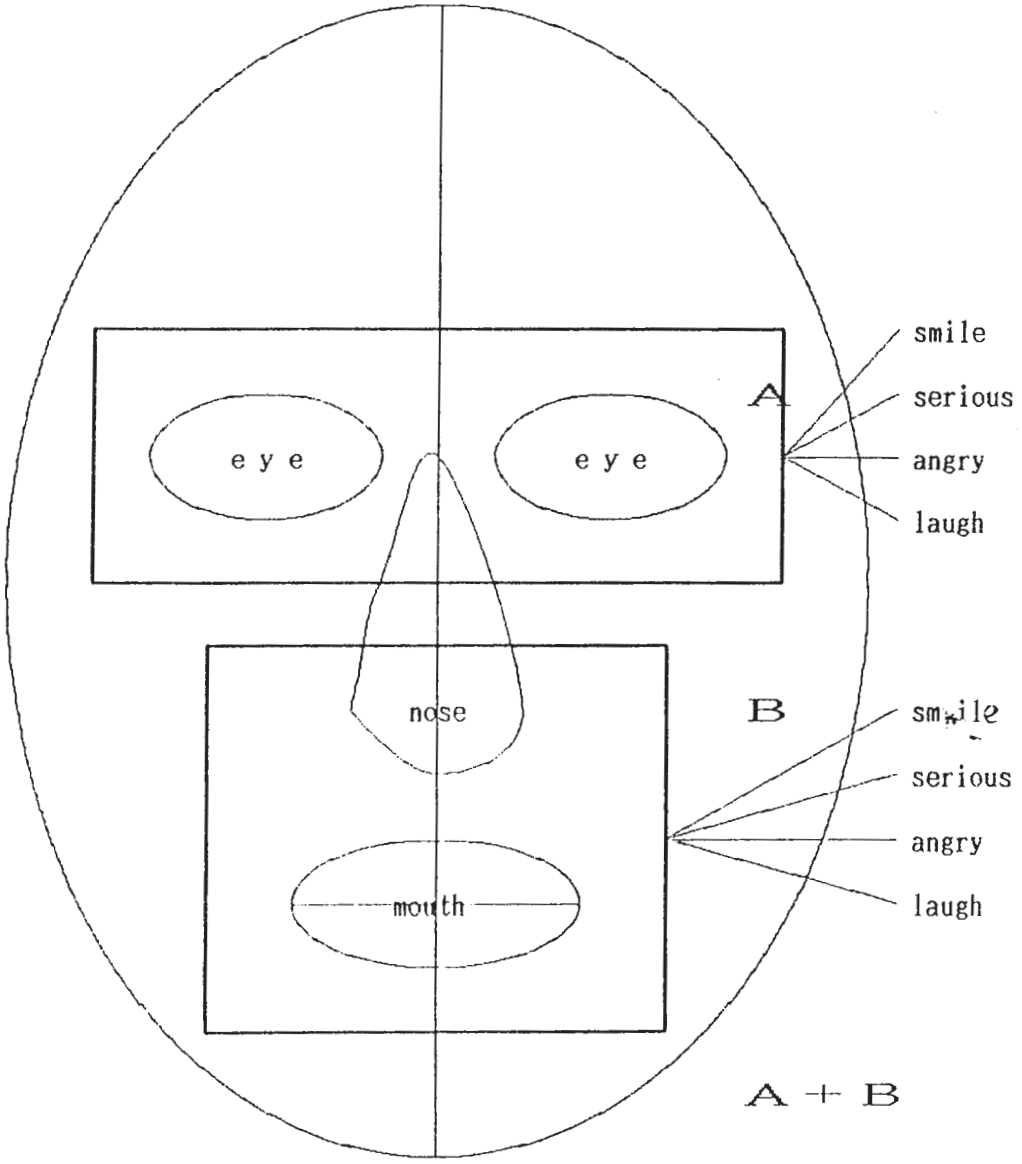


Figure 3 Decomposition into partial patterns and summation of partial matchings

の部分パターンの相対位置が少しづつ異なるため、各部分パターンの（最適な）探索が含まれ、少し複雑なシステムになる。

以上、四つの案を述べたが、一）が最も易しく、四）が最も難しい。三）と四）を実際に実験してみる価値がある。三）については、限界があることが予測されるが、どこまでできるかは一つの見所である。四）については、システムの構築により多くの努力が要るが、もっとも良い案であると思われる。

第五章 インプレメンテーションについて

今のところ、通信にパラレルマシンがないが、以上のモデルはSUNワークステーション上でも実現できる。速度の面では遅くなるが、有効性を証明するためには十分である。もちろんパラレルマシンに移植することも可能である。

現在、TVカメラから入力した画像が大概512 × 512か256 × 256画素のサイズとなっている。一画素に一つのセルをつけることが不可能なだけでなく、必要もない。画像を32 × 32画素のサイズにまず縮小しても、認識に差し支えないことが心理学の実験によっても、他のシステムの例からも、明らかになっている。そうすれば、1024個のセルの入力層になる。各画素が普通は256階調であるが、認識の場合には、適当なしきい値で二値化されれば、二値画像でもよい時がある。

案三）については、中間層は64個または16個のセルでよい。認識する表情が無表情と笑いであれば、出力層は二つのセルとなる。笑い以外の表情、例えば泣き、怒り、苦しみは出現の頻度が少ないが、十分の学習例が採れれば（しばしば困難）、対応する出力層のセルを増やせばよい。

案四）については、垂直な正面画像と仮定すれば、目と口の位置変化を垂直方向に限る。中心線を合わせることにより探索を垂直方向に限定することができる。まず目と口の各表情に対応するパターンをそれぞれのニューラル認識ネットに学習させる。出力値を0から1までの間で正規化する。そして、ラフな探索範囲の中に二つのネットを動かし、二つのネットの各表情に対応するセルの和をとる。一番大きい和を出す表情が出力される。各ネットの出力値が、ネットが正しい位置からずれると、迅速に減少するので、間違った認識結果になる確率が小さ

いと思われる。

探索の方法は、一画素づつの移動が最も簡単であるが、効率が良くない。もう一つの方法としては、まず粗いサンプルで高い出力を出す範囲を見つけ、そしてその範囲の中で細かく調べることが考えられよう。

各部分パターンの認識ネットの学習は標準的な方法で行なわれるが、どこで打ち切るかは実際の例を見て決めなければならないであろう。

結び

顔表情の認識の特徴、ニューラルなアポローチ、学習法、実現方法などについて述べた。実験をほとんどしていないが、今後どなたかの一助になれれば幸いである。

謝辞

本研究を進めるにあたって御指導下さった葉原会長、山下社長に深く感謝します。

参考文献

[1] Minsky, M. and Papert, S. (1969) Perceptrons, MIT Press.

[2] Fu. K. S. (1980) Digital Pattern Recognition, Springer-Verlag.

[3] Fu. K. S. (1982) Syntactic Pattern Recognition and Applications, Prentice-Hall.

[4] Rumelhart, D. E., McClelland, J. L. And the PDP Research Group (1986) Parallel Distributed Processing, Vol. 1, MIT Press

[5] Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986) "Learning Representations by Back-Propogating Errors," Nature, Vol. 323, pp. 533-536.

[6] 緑川、(1988) " バックプロパゲーションによる顔画像認識の一考察 "、情報処理学会第３６回全国大会資料、pp. 1881-1882

[7] 日電、読売新聞社に提出した技術紹介、（平成２年１月９日掲載）、付録参照。

[8] Fukushima, K. (1980) "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position", Biological Cybernetics, Vol. 36, pp. 193-202.

[9] 井口、宮内、(1990) “ニューラルネットを用いた顔画像認識”, 1990 年電子情報通信学会春季全国大会資料、pp. 7-232.

[10] 迫江ほか、(1988) “ダイナミックニューラルネットワークの提案ー神経回路網と DP マッチングに基づく新しい音声認識モデル”, 電子情報通信学会論文誌、Vol.J71-D, No. 7, pp.1341-1344.

日本電気株式会社

付録

ニューラルネットワークを応用した顔画像照合システムの開発に世界で初めて成功

当社はこのたび、ニューラルネットワークを用いて最大３、５００人の顔照合が可能な「顔画像照合システム」の開発に世界で初めて成功いたしました。

このシステムは、ニューラルネットワークにあらかじめ照合すべき顔を学習させることで、カメラから入力された顔と登録された顔とが一致するかどうかを瞬時に判断できる顔照合システムであります。

本システムの主な特長は、

①ＴＶカメラから入力された顔画像と登録済みの顔画像との照合を約０．１秒という高速で実行できること、

②顔画像の変化に柔軟的に対応できるニューラルネットワークアルゴリズムを用いているため、メガネを掛けたり外したりした顔でも正確に当てることができること、

③当社のデータフロー並列プロセッサ「μPD7281」を用いているため、顔学習、顔の位置合わせや入力画像の濃淡補正と拡大縮小などの前処理、照合の高速化を図ることが可能であること、

④μPD7281を64個並列にしたデータフロー並列処理装置が320MIPS という高速処理を実現しているため、１人対１００人の照合ならば０．１２秒、１人対１０００人の照合の場合は０．６秒という高速照合が可能であること、

などであります。

顔画像照合システムの仕組みは次の通りであります。

まず、画像がＲＧＢカメラから入力され、512×480のサイズにデジタル化される。次に、顔の位置合わせなどの前処理が行われ、32×32にサンプリングされた後、照合用の入力画像となる。

出力層の３層から構成されている。このネットワークを用いて、例えば、８人の中から１人を当てる場合には、入力層のデータ数は1024、中間層のデータ数は16、出力層のデータ数は８となっている。学習が終了すると、カメラの前に立った人物の顔をその場で照合することができる。また、顔照合の数は、ニューラルネットワークに新たな学習をさせるだけで自動的に増やすことができるため、カメラで顔を撮影するだけで容易に登録することができる。

システムは、６４個のデータフロープロセッサＬＳＩを並列に並べた、高速動作が可能なデータフロー並列処理システムで、パーソナルコンピュータ「ＰＣ－９８０１シリーズ」をホストコンピュータとしている。このシステムの演算部は、約30cm角の４枚のプリント板からなり、８個のプロセッサＬＳＩが１つのパイプラインリングバスに接続されているため、全部で８個のプロセッサリングを有している。各プロセッサリングは、１つのローカルメモリを有しているため、リング間での相互干渉がなく、並列演算処理が可能となっている。ニューロ演算に必要なプロセッサ間のデータ交換には、１４４本のワイアからなる高速ブロック転送バスが用いられているため、ローカルメモリと共有メモリ(8メガワード)との間で１４４ビット/５０nsの高速データ転送が可能となっている。また、システムのピーク性能は３２０ＭＩＰＳと高いため、顔画像のリアルタイム照合を実現している。本システムは、最大５１２個のＬＳＩを用いることが可能なアーキテクチャとなっているため、２５００ＭＩＰＳの高速処理を実現することも可能となっている。

本システムの主な用途としては、顔画像データベースとの高速照合が可能な類似画像検索システムへの適用、例えば、ホテルなどのＶＩＰ訪問モニタリングシステムや顔照合による自動開閉扉システムへの適用、あるいは、紙幣や切手、部品選別、医用画像診断などの外観検査システムへの適用があります。

以上

実時間顔画像照合システムの特徴

- ＴＶカメラにより顔を実時間で照合
- ニューラルネットワークによる顔画像の学習・認識機構の採用
- データフロー並列処理による高速処理の実現

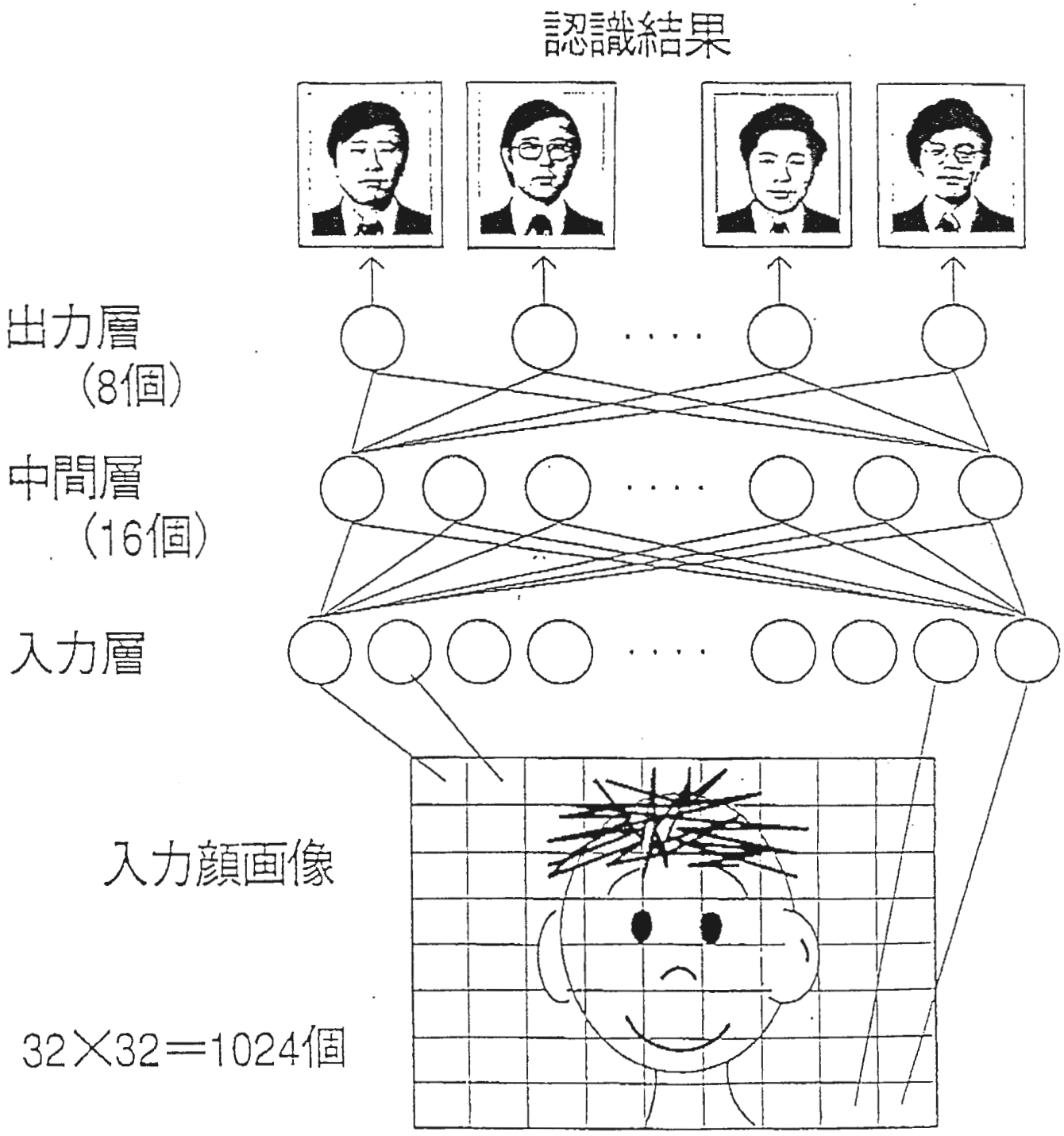
  64 並列（最大512まで拡張可）
  320MOPS（最大2500MOPS）

- 照合時間

プロトタイプ（64並列）	最大字構成時（512並列）
100人 0.116秒	0.042秒
500人 0.280秒	0.064秒
1,000人 0.594秒	0.104秒
5,000人 7.620秒	0.990秒

ニューラルネットワーク構造

- 入力層、中間層、出力層の３層構造
- バックプロパゲーションで学習



第三部

線画における一般円柱の解釈

Research Note submitted to the International Journal of ARTIFICIAL INTELLIGENCE

Beyond commenting on Horaud and Brady's paper "On the geometric interpretation of image contours"

Gang XU and Hiromi T. TANAKA
ATR Communication Systems Research Laboratories
Seika-cho, Soraku-gun, Kyoto 619-02, Japan

Beyond commenting on Horaud and Brady's paper

Abstract

Horaud and Brady [1] have recently proposed a computational model for the 3D interpretation of image contours. While we believe that the idea of combining constraints from extremal contours with constraints from discontinuity contours is significant, we point out: (1) that the slant-tilt is a better representation for space orientation than the Gaussian sphere; and (2) that the compactness measure M should only be employed when no other sources of information are available. We first show that using the slant-tilt representation reduces the particular example given in [1] from a 2D problem to a 1D problem, then present algorithms for determining the tilt of the cylinder axis for a broader class of Straight Homogeneous Generalized Cylinders (rather than only circular generalized cylinders), and finally propose algorithms for determining the slant of the cylinder axis.

1. Introduction

In their recent paper [1], Horaud and Brady proposed the idea of combining constraints from extremal contours with constraints from discontinuity contours. They first classify image contours into extremal contours and discontinuity contours, and then combine constraints from each contour to form unified interpretations of the surfaces. The specific example they employ to show their idea is a generalized cylinder with a straight axis and a circular cross section which is perpendicular to the straight axis. The medium in which the constraints are combined is the Gaussian sphere. In Section 2, we first briefly introduce the Gaussian sphere and the slant-tilt system as representations of space orientation, and then point out that representing the surface normal by slant and tilt rather than α and β in the Gaussian sphere is more clear and reduces the problem from 2-dimensionality to 1-dimensionality in the specific example given in [1]. In section 3, we present algorithms for determining the tilt of the cylinder axis for a broader class of Straight Homogeneous Generalized Cylinders (SHGC's, rather than only circular generalized cylinders). In Section 4, we propose algorithms for determining the slant of the cylinder axis, for which

maximizing M, as done in [1], does not necessarily provide correct solutions.

2. α-β vs. Slant-tilt

The Gaussian sphere and the slant-tilt system are two representations for space orientation. The Gaussian sphere is a sphere of unit radius. A space vector is mapped to a point on the sphere, sharing the same orientation with the line radiating from the origin to that point. The point on the sphere is represented in the spherical coordinate system by two angles, α and β (Fig. 1). Its associated space vector is $(\cos\alpha\,\cos\beta, \sin\alpha\,\cos\beta, \sin\beta)$. The slant-tilt system is defined with respect to the viewing direction (Fig. 2). Slant σ is the angle between the viewing direction and the surface normal, while tilt τ is the angle between the horizontal coordinate axis and the normal's projection onto the image plane [2,3]. In vector form, the orientation is $(\sin\sigma\,\cos\tau, \sin\sigma\,\sin\tau, \cos\tau)$. Throughout this paper, as in Horaud and Brady's original paper, we assume orthographic projection.

In the specific case of a circular generalized cylinder, as explained in detail in [1], the symmetry axis of the pair of extremal boundaries is the image of the cylinder axis. Intuitively, the 3D orientation of the axis lies on the plane defined by the symmetry axis and the viewing direction, leaving only one degree of freedom to be determined. Describing the constraint in the Gaussian sphere in terms of α and β, we have an S-shaped curve.

On the other hand, Brady and Yuille [5] have defined a compactness measure M for a closed planar contour, which is a function of the 3D orientation of the plane in which the contour lies, as the ratio of the enclosed area to the perimeter squared. By maximizing this measure, one can determine the most symmetrical and most compact interpretation of that contour. For an image ellipse, maximizing M results in the interpretation of a circle, similar to the result of human perception. Again, describing M in terms of α and β, we have a surface. For an image ellipse, the surface is a saddle-like surface with two peaks of the same height which correspond to the two space circles, one being a Necker reversal of the other.

Combining the above two constraints in the Gaussian sphere is, as described in [1], to overlap the S-shaped curve with the saddle-like surface. Specifying M at the maximums or any other values, we can determine pairs of α and β. If M's two maxima do not lie along the S-shaped curve, then the interpretation is that the cross-section is not perpendicular to the cylinder axis.

In the following, we interpret the same figure using slant-tilt to represent the orientation. First, since the symmetry axis is the projection of the cylinder axis, tilt can be directly determined as the angle the symmetry axis makes with the horizontal axis of the image plane (because it can take two directions, there are two possible values), leaving only slant to be determined. On the other hand, since the figure is interpreted in terms of a circular generalied cylinder, the cross section should always be projected as an ellipse. An ellipse has a major axis a and a minor axis b. Only if the minor axis coincides with the symmetry axis of the pair of the extremal contours, is it possible for the cross section to be perpendicular to the cylinder axis. Otherwise, the figure cannot be interpreted as a generalized cylinder with circular cross sections perpendicular to the straight axis. Now the problem reduces to one of determining slant. One way is to maximize M. Fig. 3 qualitatively shows M as a function of slant for the ellipse used in [1]. At zero slant, M is calculated as the ratio of the image area over the image perimeter squared. M reaches its maximum as slant approaches $cos^{-1}(\frac{b}{a})$. This corresponds to the interpretation of a surface of revolution with the circular cross sections perpendicular to the cylinder axis, which is slanted by $cos^{-1}(\frac{b}{a})$ around the major axis. M approaches zero as slant approaches 90°.

3. Determining the tilt of the cylinder axis

Horaud and Brady restrict their analysis [1] to surfaces of revolution, which are the only case where there is always a symmetry axis of the extremal contours. For a more general cylinder, e.g., a Straight Homogeneous Generalized Cylinder (SHGC), as in the wide accepted definition [4,5], there generally does not exist any symmetry in the extremaml contours. Thus, we need a more general algorithm to determine the image of the cylinder axis, i.e., the tilt of the cylinder axis.

In their recent work, Ponce *et al.* [4] have proposed two algorithms to determine the axis of a straight homogeneous generalized cylinder (SHGC) in image. The first algorithm is a Hough transformation algorithm which uses the property that tangents to the extremal contours at points which belong to the same cross section intersect on the image of the axis. The second algorithm is also based on the above property, but restricts the points to be those zeros of curvature along the extremal contours from the property that if the sweeping rule curve of an SHGC has zeros of curvature, all points on the same cross section are zeros of curvature. While the first algorithm is always applicable, the second one is not necessarily so because there are not necessarily zeros of curvature along the extremal contours.

Since here we have a discontinuity contour which is to be interpreted as a cross section, there are two intersection points between the discontinuity contour and the extremal contours. It is known that the tangents to the extremal contours at the two points intersect on the image of the cylinder axis and determine one point on it. Using this information reduces the Hough space in the first algorithm from two dimensions to one dimension.

In many cases, we can determine another point on the cylinder axis by finding the centroid of the area that the discontinuity contour encloses, from the fact that an SHGC is usually produced by translating and scaling a planar cross section along a straight axis through the centroid of the cross section. We call this SHGC a Straight Centric Homogeneous Generalized Cylinders (SCHGC, Fig. 4). The additional assumption is that

the straight axis passes through the centroids of the cross sections. It is a reasonable assumption because this is the most stable structure and most man-made generalized cylinders satisfy this assumption.

In the above algorithm, a property is implicitly employed that *the centroid of the region that a discontinuity contour encloses is the projection of the centroid of the cross section that the contour depicts.*

Proof: The ratio of two areas on the same plane is view-invariant under the orthographic projection. In space, any lines which pass through the centroid of a planar region enclosed by a boundary divide the region into equal halves (of the same area). In image, the projections of the lines also divide the region enclosed by the projected boundary into equal halves. therefore, the centroid of the region in image, as the intersection of the projected lines, is the projection of the centroid of the planar region in space.

If the discontinuity contour is skewed symmetrical, then the centroid is on the skewed symmetry axis. If the cross section is centro-symmetrical, then from the tangent invariance property, the centroid is the intersection of two lines that respectively links points on the discontinuity contour which have the same tangents to the contour.

Whether or not an SHGC is an SCHGC can be examined by looking at another point on the axis, e.g., a centroid associated with another discontinuity contour, or, an intersection point of two tangents to the extremal contours at points on another cross section. If the three points lie on the same line, then highly probably the SHGC is an SCHGC. If they do not, then the SHGC is not an SCHGC, and the axis cannot be determined by finding the centroid associated with the discontinuity contour.

4. Determining the slant of the cylinder axis

In this section, we point out that the approach of maximizing the compactness measure M should only be taken when no other sources of information are available, and propose alternative ways to determine the slant of the cylinder axis.

First, we show that maximizing M usually does not provide correct solutions. Take, for example, a parallelogram (Fig. 5). Maximizing M allows interpreting the parallelogram as a square in space, determining a unique 3D orientation with a Necker reversal for the square, which disagrees with the tilt of the cylinder axis obtained in the last section. To put it another way, maximizing M plus the tilt of the cylinder axis overconstrains the interpretation. In fact, this figure should be interpreted as a generalized cylinder with a rectangular (not square) cross section perpendicular to the axis.

If there is any knowledge about the cross section, then it is a better way to use it. For example, if the discontinuity contour is an ellipse whose minor axis coincides with the tilt of the cylinder axis, then we always interpret the ellipse as a circle in space. The slant can be directly determined as the arc cosine of the minor axis over the major axis.

Another example is to use symmetry information in the cross section. If the discontinuity contour is skewed symmetrical, then it is reasonable to infer that the original cross section is symetrical [6]. For a skewed symmetrical closed contour, there are two directions: the skewed symmetry direction and the skewed transverse direction. While they are perpendicular in space, they do not appear so in image (Fig. 6). It is straightforward to use this constraint to determine the slant of the cylinder axis [7,8]. Here, we have three tilts: the tilts of the symmetry axis and the transverse axis (always through the centroid) and the tilt of the cylinder axis. Assuming the cross section is perpendicular to the cylinder axis, we have three equations describing that the inner products of every two space vectors are zero. The slant of the cylinder axis σ_c can then be determined as

$$\sigma_c = \sin^{-1}\sqrt{-\frac{\cos(\tau_s - \tau_t)}{\cos(\tau_s - \tau_c)\cos(\tau_t - \tau_c)}}, \tag{1}$$

the slant of the symmetry axis as

$$\sigma_s = \sin^{-1}\sqrt{-\frac{\cos(\tau_t - \tau_c)}{\cos(\tau_t - \tau_s)\cos(\tau_c - \tau_s)}}, \tag{2}$$

and the slant of the transverse axis as

$$\sigma_t = \sin^{-1}\sqrt{-\frac{\cos(\tau_c - \tau_s)}{\cos(\tau_c - \tau_t)\cos(\tau_s - \tau_t)}}, \tag{3}$$

where τ_s, τ_t and τ_c are the tilts of the symmetry axis, the transverse axis and the cylinder axis, respectively.

As evident from the equation, for σ_c to have a solution, τ_s, τ_t and τ_c must satisfy the following condition: the angle between the skewed symmetry axis and the skewed transverse axis is smaller than 180° and larger than 90° (if smaller than 90°, then, without loss of generality, its supplementary angle is used), and both the angle between the tilt of the cylinder axis and the skewed symmetry axis as well as the angle between the tilt of the cylinder axis and the skewed transverse axis are smaller than 90°. If the three tilts do not satisfy this condition, then the cross section cannot be interpreted as perpendicular to the cylinder axis. A counterexample is shown in Fig. 7, which is reproduced by turning the parallelogram cross section in Fig. 5 90°. Since the three tilts do not satisfy the above condition, the figure can no longer be interpreted as a rectangular block.

5. Conclusion

Horaud and Brady have recently proposed the idea of combining constraints from extremal contours and constraints from discontinuity contours in [1]. However, as we have shown in Section 2, the use of the Gaussian sphere to represent surface orientation complicates computation. As a more intuitive and direct representation of surface normal, the slant-tilt system reduces the particular problem of interpreting a circular generalized cylinder from 2-dimensional to 1-dimensional. In the following sections, we have expanded the analysis from circular generalized cylinders to straight homogeneous generalized cylinders and proposed algorithms to determine the tilts and slants of the axes of the SHGC's. We have also discussed that maximizing M, as originally proposed, does not necessarily produce correct solutions and thus should only be employed when no other sources of information are available. As an alternative, we have shown how to interpret a skewed symmetrical discontinuity contour as a symmetrical cross section and to determine the orientation of the cylinder axis.

Acknowledgment

The authors would like to thank Kohei Habara, Chairman of the ATR Governing Board, and Koichi Yamashita, President of ATR Communication Systems Research Laboratories, for their advice and encouragement.

References

1. Horaud, R. and Brady, M., On the geometric interpretation of image contours, *Artificial Intelligence* **37**, (1988) 333-353

2. Witkin, A., Recovering surface shape and orientation from texture, *Artificial Intelligence* **17**, (1981) 17-45

3. Stevens, K., Slant-tilt: The visual encoding of surface orientation, *Biological Cybernetics* **46**, (1983) 183-195

4. Ponce, J., Chelberg,D., and Mann, W., Invariant Properties of Straight Homogeneous Generalized Cylinders and Their Contours, *IEEE Tran. on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 9, September 1989

5. Shafer, S., *Shadows and Silhouettes in Computer Vision,* New York: Kluwer Academic, (1985)

6. Kanade, T. Recovery of the three-dimensional shape of an object from a single view, *Artificial Intelligence* **17**, (1981) 409-460

7. Stevens, K., The visual interpretation of surface contours, *Artificial Intelligence* **17**, (1981) 47-73

8. Xu, G. and Tsuji, S., Inferring surfaces from boundaries, in: *Proc. of 1st International Conference on Computer Vision*, (1987) 716-720

Fig. 1 The Gaussian sphere representation

Fig. 2 The slant-tilt representation

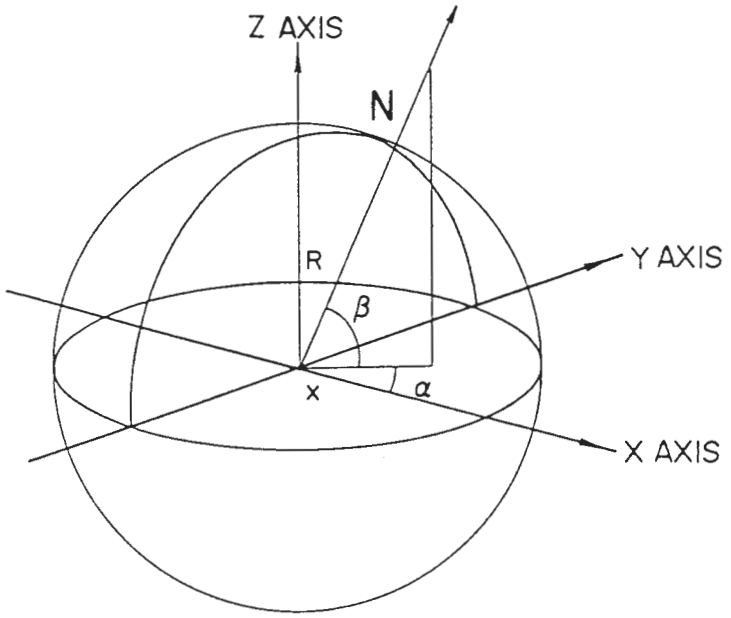
Fig. 3 The curve qualitatively illustrates the compactness measure M as a function of the slant for the ellipse in [1].

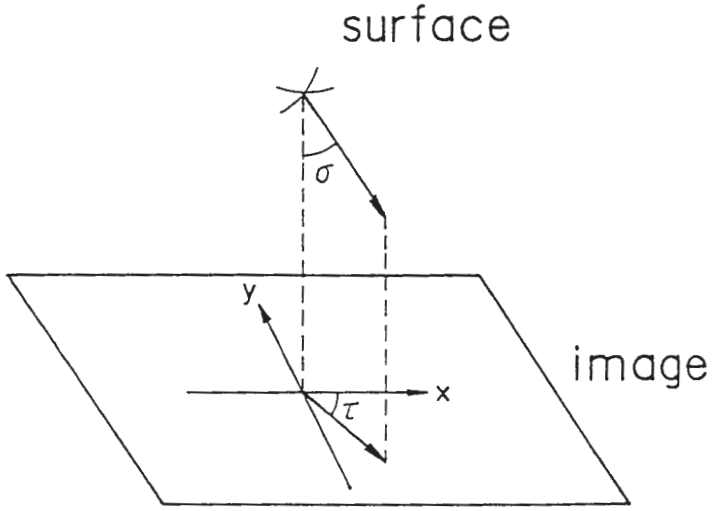
Fig. 4 A straight homogeneous generalized cylinder with the axis through the centroids of the cross sections. The cylinder axis can be determined by finding the intersection of the tangents to the extremal contours at their two intersections with the discontinuity contour, and the centroid of the area enclosed by the discontinuity contour.

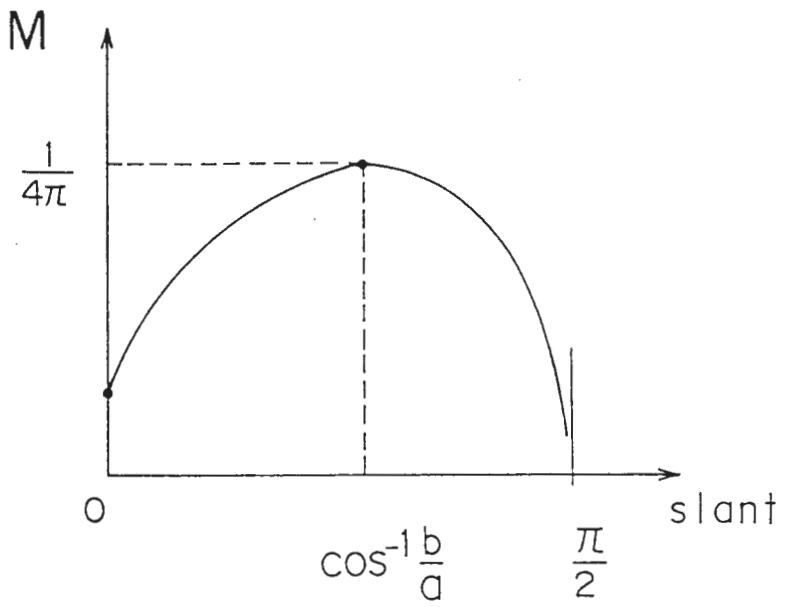
Fig. 5 Maximizing M interprets the parallelogram as a rectangle, which makewould it impossible to interpret the cross section as being perpendicular to the generalized cylinder axis.

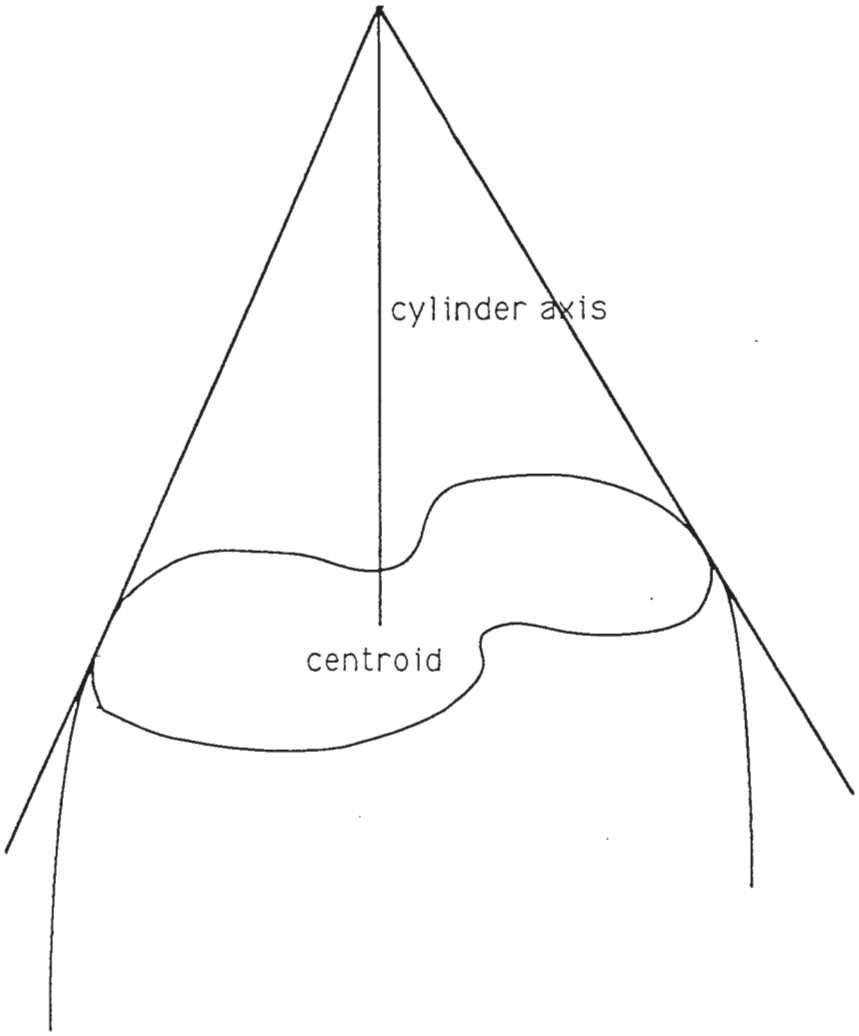
Fig. 6 The skewed symmetry discontinuity contour, with a skewed symmetry axis and a skewed transverse axis, is interpreted as a symmetrical cross section.

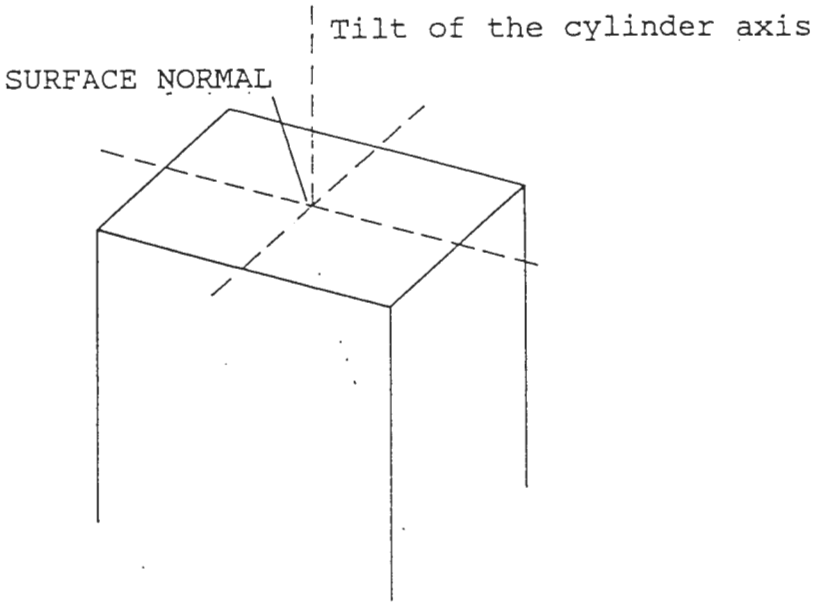
Fig. 7 Turning the parallelogram cross section in Fig. 5 90°, the figure can no longer be interpreted as a rectangular block.

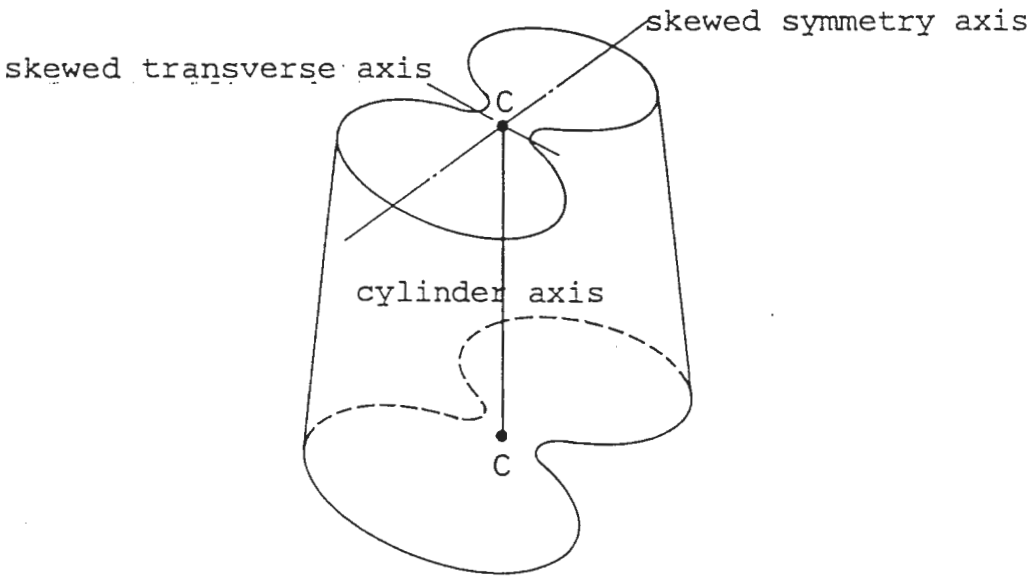

Z AXIS
N
R
Y AXIS
β
x
α
X AXIS


surface
σ
y
x
τ
image


M
$\frac{1}{4\pi}$
0
slant
$\cos^{-1}\frac{b}{a}$
$\frac{\pi}{2}$


cylinder axis
centroid


Tilt of the cylinder axis
SURFACE NORMAL


skewed symmetry axis
skewed transverse axis
C
cylinder axis
C