Deformable Templates for Face Recognition

Alan L. Yuille
Division of Applied Science
Harvard University

Abstract

■ We describe an approach for extracting facial features from images and for determining the spatial organization between these features using the concept of a deformable template. This is a parameterized geometric model of the object to be recognized together with a measure of how well it fits the image data. Variations in the parameters correspond to allowable deformations of the object and can be specified by a probabilistic model. After the extraction stage the parameters of the deformable template can be used for object description and recognition. ■

INTRODUCTION

The sophistication of the human visual system is often taken for granted until we try to design artificial systems with similar capabilities. In particular humans have an amazing ability to recognize faces seen from different viewpoints, with various expressions and under a variety of lighting conditions. It is hoped that current attempts to build computer vision face recognition systems will shed light on how humans perform this task and the difficulties they overcome.

In this article we describe one promising approach toward building a face recognition system. Some alternative approaches are reviewed in Turk and Pentland (1989).

A standard way of describing a face consists of representing the features and the spatial relations between them. Indeed two of the earliest face recognition systems (Goldstein, Harmon, & Lesk, 1972; Kanade, 1977) used features and spatial relations, respectively. Such representations are independent of lighting conditions, may be able to characterize different spatial expressions, and have some limited viewpoint invariance.

These representations, however, are extremely difficult to compute reliably from a photograph of a face. It is very hard to extract features, such as the eyes, by using current computer vision techniques. Deformable templates, however, offer a promising way of using a priori knowledge about features, and the spatial relationships between them, in the detection stage. Once the feature has been extracted the parameters of the deformable template can be used for description and recognition.

Designing a deformable template to detect a feature, or an object, falls into two parts: (1) providing a geometric model for the template, and (2) specifying a model, the *imaging model*, for how a template of specific geometry will appear in the image and a corresponding measure of fitness, or matching criterion, to determine how the template interacts with the image.

For many objects it is straightforward to specify a plausible geometric model. Intuition is often a useful guide and one can draw on the experience of artists (for example, Bridgman, 1973). Once the form of the model is specified it can be evaluated and the probability distribution of the parameters determined by statistical tests, given enough instances of the object.

Specifying the imaging model and the matching criterion is often considerably harder. The way the object reflects light depends both on the reflectance function of the object, which in principle can be modeled, and on the scene illumination, which is usually unknown. The matching criterion, however, should aim to be relatively independent of the lighting conditions. An even more serious problem arises when part of the object is invisible, perhaps due to occlusion by another object or by lying in deep shadow. Most matching criteria will break down in this situation but recent work, described in the fourth section, shows promise of dealing with such problems.

The plan of this chapter is as follows. The second section gives a simple introduction to deformable templates by describing the pioneering work of Fischler and Elschlager (1973) on extracting global descriptions of faces. The third section gives a more detailed description of using deformable templates to extract facial features (Yuille, Cohen, Hallinan, 1989). The fourth section shows a way to put deformable templates in a more robust framework (Hallinan & Mumford, 1990), which promises to make them more reliable and effective.

Deformable templates, and the closely related elastic models (Burr, 1981a,b; Durbin & Willshaw, 1987; Durbin, Szeliski, & Yuille, 1989) and snakes (Kass, Witkin, &

Terzopolous, 1987; Terzopolous, Witkin, & Kass, 1987), have been extensively used in recent years in vision and related areas. Much of this work fits into the general Bayesian statistical framework developed by Grenander and his collaborators (Chow, Grenander, & Keenan, 1989; Grenander, 1989; Knoerr, 1989) for synthesizing and recognizing biological shapes.

GLOBAL TEMPLATES

In Fischler and Elschlager's approach (1973) the face is modeled by a set of basic features connected by springs (see Fig. 1). The set of features includes the eyes, hair, mouth, nose, and left and right edges. The individual features do *not* deform and each feature has a local measure of fit to the image. During the matching stage the entire structure is deformed, rather like a rubber sheet, until all features have a good local fit and the spring forces are balanced. The springs help ensure that the spatial relations between the features are reasonable (i.e., the nose is not above the hair).

More specifically, this approach defines a local fitness measure $l_i(x_i)$ that indicates how strongly the ith feature fits at location x_i. Fischler and Elschlager define simple fitness measures for each feature. For example, for the left edge of the face the fitness measure at x_i is the difference between the sums of the four intensity values to the left and right of x_i. Thus this measure is large for a straight vertical line with low intensity on the left and high intensity on the right.

The spring joining the ith and jth features are given a cost function $g_{ij}(x_i, x_j)$, where x_i and x_j are the positions of the features. In most cases g_{ij} is assumed to be symmetric and to depend only on the relative positions of the features, hence it can be written as $g_{ij}(x_i - x_j)$. Not all features are joined by springs (see Fig. 1), so for each feature i we let N_i denote the set of features to which it is connected.

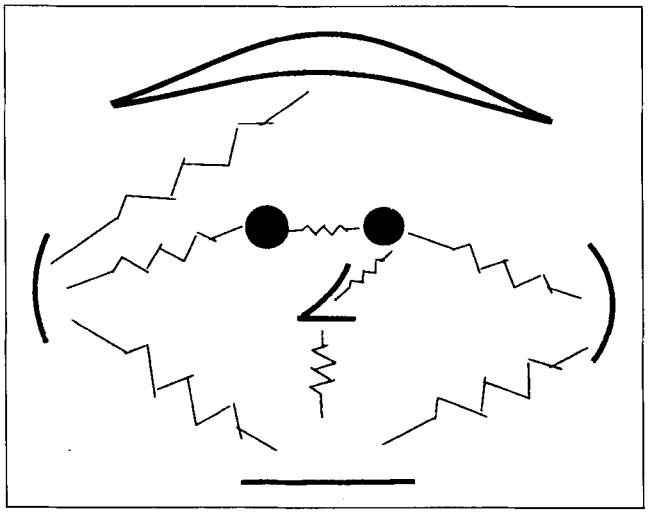

Figure 1. The face model consists of the hair, eyes, nose, mouth, and the left and right edges joined together by springs.

The total measure of fitness is the sum of the local fitness measures for each feature and the cost of the spring terms. Let $X_7 = \{x_1, x_2, \ldots, x_7\}$ be the positions of the features. The fitness is

$$E(X_7) = \sum_{i=1}^{7} l_i(x_i) + \sum_{i=1}^{7} \left\{ \sum_{j \in N_i} g_{ij}(x_i - x_j) \right\} \qquad (1)$$

The theory suggests localizing the features in a specific face by adjusting X_7 to maximize $E(X_7)$. This is a hard combinatoric problem and Fischler and Elschlager suggest an algorithm, the *linear embedding algorithm*, to solve it. We will not, however, be concerned here with the details of the algorithm, which are related to dynamic programming.

The system was shown to work on a number of 35 by 40 images. The authors reported that the main errors occurred in the location of the mouth/nose complex, although the other features were correctly located. In these situations the spring forces put the mouth/nose complex in roughly the right position but were unable to localize it correctly. They also mentioned that in the presence of noise the simple local fitness measures seemed inadequate to accurately locate the nose and mouth.

Computer simulations at the Harvard Robotics Lab (U. Wehmeier, personal communication) show that closely related techniques work well for images of faces with constant size and with little noise. In these simulations five features were chosen: (1) the eyes, (2) the nostrils of the nose, and (3) the sides of the mouth. These features were attracted to valleys, dark blobs, in the image intensity and were connected by springs. A steepest descent algorithm was used (see Fig. 2). Once the positions of the features were located approximately, then more accurate detectors can be used to find them more precisely.

Fischler and Elschlager's approach is very attractive but it has several weaknesses in its present form. The main problem is the simplicity of the local fitness measures. At present they are strongly scale dependent and may fail under certain noise conditions. The spring forces might also be improved, by a systematic study of the spatial relations, to make sure they give a more accurate bias.

Systems of this type seem to work best on small pictures with coarse levels of resolution. At coarse scales the local fitness measures may be very simple (Wehmeier was able to locate many features in terms of valleys). At finer scales, however, the local fitness measures must become more complex to take into account the greater variability of the feature. This suggests a coarse to fine, or pyramidal (Burt & Adelson, 1983), approach in which a global template is used to identify likely positions in the coarse image that can be located and (we hope) verified by more sophisticated techniques at a finer scale.

We will now describe a system that uses deformable templates to locate the features themselves.

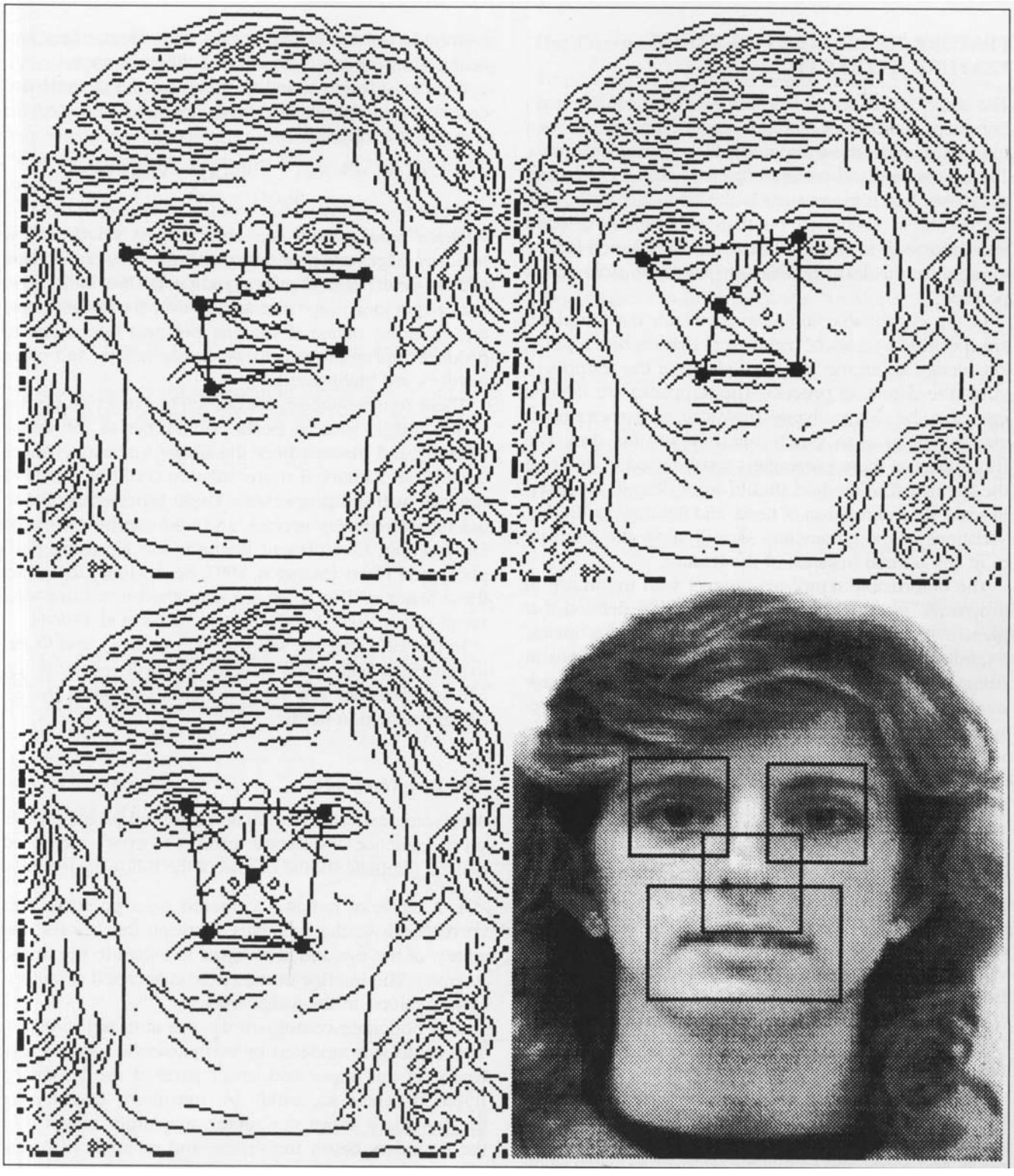

Figure 2. A time sequence showing how the estimated positions of the five features, represented by black dots, evolve as they are attracted to valleys in the intensity under the influence of the springs, represented by thick lines. For reasons of clarity this evolution is shown on an edge image of the face, rather than on the face itself. The dots locate the features approximately and specify regions of interest within which more sophisticated detectors can locate the features more accurately. Figure courtesy of U. Wehmeier.

FEATURE TEMPLATES FOR FACIAL FEATURE EXTRACTION

The ability to detect and describe salient features is an important component of a face recognition system. Such features include the eyes, nose, mouth, and eyebrows. This task is hard and current edge detectors seem unable to reliably find features such as the boundary of the eye. The problem seems to be that, although it is often straightforward to find local evidence for edges, it is hard to organize this local information into a sensible global percept.

In the deformable template approach the templates are specified by a set of parameters that enables a priori knowledge about the expected shape of the features to guide the detection process. The templates are flexible enough to be able to change their size, and other parameter values, so as to match themselves to the data. The final values of these parameters can be used to describe the features. The method should work despite variations in scale, tilt and rotation of head, and lighting conditions. Variations of the parameters should allow the template to fit any normal instance of the feature.

The deformable templates interact with the image in a dynamic manner. An energy function is defined that gives a measure of fit of the template to the image. Minimizing the energy attracts the template to salient features, such as peaks, valleys, and edges in the image intensity. The minimum of the energy function corresponds to the best (local) fit with the image. The template is given some initial parameters that are then updated by steepest descent. This corresponds to following a path in parameter space, and contrasts with traditional methods of template matching that would involve sampling the parameter space to find the best match (and whose computational cost increases exponentially with the dimension of the parameter space). Changing these parameters corresponds to altering the position, orientation, size, and other properties of the template. The initial values of the parameters, which may be very different from the final values, are determined by preprocessing. If, for example, we have input from a global face template (as described in the previous section) then we could use this input to determine likely initial values.

The template is designed to act on representations of the image, as well as on the image itself. These representations are based on fields, which highlight valleys, peaks, and edges, and enable the template to match when its initial parameter values are very different from the correct ones. The final fitness measure, however, is mostly independent of these representations.

Representations of the Image

We preprocess the image to obtain fields, $\Phi_v(\mathbf{x})$, $\Phi_e(\mathbf{x})$, and $\Phi_p(\mathbf{x})$, (representing valleys, edges, and peaks) on which the deformable template will act. These fields are intended to take their largest values at valleys, edges, and peaks. They are calculated in two different ways.

For final precision, and to minimize the dependence on prior assumptions, the fields are defined directly in terms of the intensity,

$$\Phi_v(\mathbf{x}) = -I(\mathbf{x}), \qquad \Phi_e(\mathbf{x}) = \nabla I(\mathbf{x})\cdot\nabla I(\mathbf{x}), \qquad \Phi_p(\mathbf{x}) = I(\mathbf{x}) \tag{2}$$

These fields clearly take their largest values at low intensity, edges and peaks, respectively. They can be used to help detect valleys, edges, and peaks. It is hard, however, to get long range interactions over the image. These are easier to obtain if we first perform operations to produce representations in which the valleys, and other features, are highlighted.

These representations are chosen to extract properties of the image, such as peaks and valleys in the image intensity and places where the image intensity changes quickly (an additional representation could be added to describe textural properties). These representations do not have to be very precise, and they can be calculated fairly simply. Our present methods involve using morphological filters (Maragos, 1987; Serra, 1982) to extract these features. The fields are smoothed to ensure long range interactions; for details see Yuille et al. (1988).

In the following we will use $\Phi_v(\mathbf{x})$, $\Phi_e(\mathbf{x})$, and $\Phi_p(\mathbf{x})$ to represent both types of fields and will specify in the text whether they are intensity fields, calculated by (2), or representation fields.

The Eye Template

After some experimentation and informal psychophysics on the salience of different features of eyes we decided that the template should consist of the following features:

1. A circle of radius r, centered on a point $\mathbf{x}_c$. This corresponds to the boundary between the iris and the whites of the eye and is attracted to edges in the image intensity. The interior of the circle is attracted to valleys, or low values, in the image intensity.

2. A bounding contour of the eye attracted to edges. This contour is modeled by two parabolic sections representing the upper and lower parts of the boundary. It has a center $\mathbf{x}_e$, width $2b$, maximum height a of the boundary above the center, maximum height c of the boundary below the center, and an angle of orientation θ.

3. Two points, corresponding to the centers of the whites of the eyes, which are attracted to peaks in the image intensity. These points are labeled by $\mathbf{x}_e + p_1(\cos\theta, \sin\theta)$ and $\mathbf{x}_e + p_2(\cos\theta, \sin\theta)$, where $p_1 \geq 0$ and $p_2 \leq 0$. The point $\mathbf{x}_e$ lies at the center of the eye and θ corresponds to the orientation of the eye.

4. The regions between the bounding contour and the iris also correspond to the whites of the eyes. They will be attracted to large values in the image intensity.

These components are linked together by three types of forces: (1) forces that encourage $\mathbf{x}_c$ and $\mathbf{x}_e$ to be close together, (2) forces that make the width $2b$ of the eye roughly four times the radius r of the iris, and (3) forces that encourage the centers of the whites of the eyes to be roughly midway from the center of the eye to the boundary.

The template is illustrated in Figure 3. It has a total of 9 parameters: $\mathbf{x}_c$, $\mathbf{x}_e$, p_1, p_2, r, a, b, c, and θ. All of these are allowed to vary during the matching.

To give the explicit representation for the boundary we first define two unit vectors

$$\mathbf{e}_1 = (\cos\theta, \sin\theta), \qquad \mathbf{e}_2 = (-\sin\theta, \cos\theta) \quad (3)$$

which change as the orientation of the eye changes. A point $\mathbf{x}$ in space can be represented by (x_1, x_2) where

$$\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2. \quad (4)$$

Using these coordinates the top half of the boundary can be represented by a section of a parabola with $x_1 \in [-b, b]$

$$x_2 = a - \frac{a}{b^2}x_1^2 \quad (5)$$

Note that the maximal height, x_2, of the parabola is a and the height is zero at $x_1 = \pm b$. Similarly the lower half of the boundary is given by

$$x_2 = -c + \frac{c}{b^2}x_1^2 \quad (6)$$

where $x_1 \in [-b, b]$.

The Energy Function for the Eye Template

We now define a potential energy function for the image that will be minimized as a function of the parameters of the template. This energy function not only ensures that the algorithm will converge, by acting as a Lyaponov function, but also gives a measure of the goodness of fit of the template. More robust energy functions are described in section four.

The complete energy function $E_c(\mathbf{x}_e, \mathbf{x}_c, p_1, p_2, a, b, c, r, \theta)$ is given as a combination of terms due to valley, edge, peak, image, and internal potentials. More precisely,

$$E_c = E_v + E_e + E_i + E_p + E_{\text{internal}} \quad (7)$$

where

1. The intensity/representation valley potentials are given by the integral of the intensity/representation fields over the interior of the circle divided by the area of the circle,

$$E_v = -\frac{c_1}{Area}\int\int_{\text{Circle-Area}} \Phi_v(\mathbf{x})\, dA \quad (8)$$

2. The intensity/representation edge potentials are given by the integrals of the intensity/representation edge fields over the boundaries of the circle divided by its length and over the parabolas divided by their lengths,

$$E_e = -\frac{c_2}{Length}\int_{\text{Circle-Bound}} \Phi_e(\mathbf{x})\, ds - \frac{c_3}{Length}\int_{\text{Para-Bound}} \Phi_e(\mathbf{x})\, ds \quad (9)$$

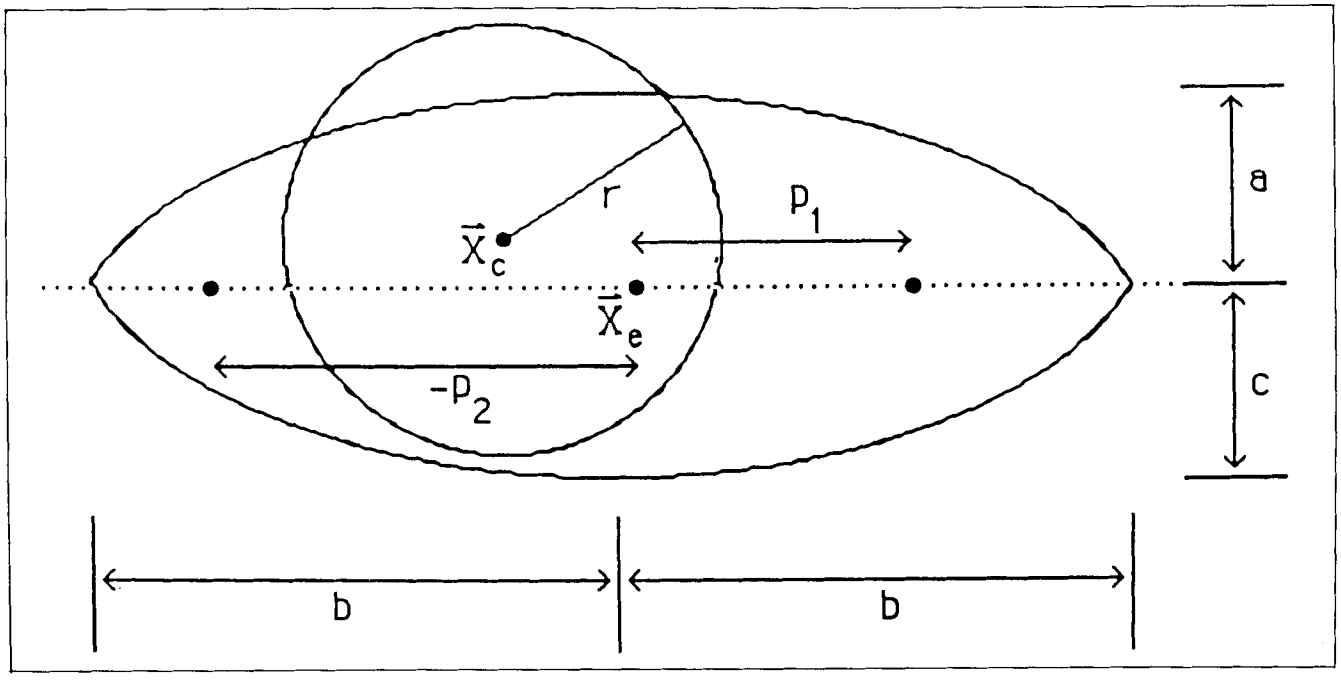

Figure 3. The eye template.

3. The intensity peak potentials attempt to maximize the intensity peak field, $\Phi_p(\mathbf{x})$, between the circle and the parabolas, again divided by the area,

$$E_p = -\frac{c_5}{Area}\int \int_{Whites} \Phi_p(\mathbf{x})\, dA \qquad (10)$$

4. The representation peak potentials are the representation peak field evaluated at the two peak points and are given by

$$E_p = c_6\{\Phi(\mathbf{x}_e + p_1\mathbf{e}_1) + \Phi(\mathbf{x}_e + p_2\mathbf{e}_1)\} \qquad (11)$$

5. The internal potentials are given by

$$E_{internal} = \frac{k_1}{2}(\mathbf{x}_e - \mathbf{x}_c)^2 + \frac{k_2}{2}(p_1 - \frac{1}{2}\{r + b\})^2 +$$

$$\frac{k_2}{2}(p_2 + \frac{1}{2}\{r + b\})^2 + \frac{k_3}{2}(b - 2r)^2. \qquad (12)$$

The $\{c_i\}$ and $\{k_i\}$ are usually fixed coefficients but we will allow them to change values (corresponding to different epochs) as the process proceeds. Changing the values of these coefficients enables us to use a matching strategy in which different parts of the template guide the matching at different stages. For example, the valley in the image intensity corresponding to the iris is very salient and is more effective at "attracting" the template from long distances than any other feature. Thus its strength, which is proportional to c_1, should initially be large. Orienting the template correctly is usually best performed by the peak terms, thus c_6 should be large in the middle period. The constants c_2 and c_3 can then be increased to help find the edges. Finally, the terms involving the image intensity can be used to make fine scale corrections and determine the final measure of fit. This corresponds to a strategy in which the position of the eye is mainly found by the valley force, the orientation by the peak force, and the fine scale detail by the edge and intensity forces. In this scenario the values of the cs will be changed dynamically. Typical values for the coefficients are

$$(c_1,c_2,c_3,c_4,c_5,c_6) \approx (4000,50,50,125,150,50)$$

and

$$(k_1,k_2,k_3) \approx (10,1,0.05).$$

The individual energy terms can be written as functions of the parameter values. For example, the sum over the boundary can be expressed as an integral function of $\mathbf{x}_e$, a, b, c, and θ by

$$\int_{Para-Bound} \Phi_e(\mathbf{x})\, ds$$

$$= \frac{c_3}{Length}\int_{x_1=-b}^{x_2=b} \Phi_e(\mathbf{x}_e + x_1\mathbf{e}_1 + \left\{a - \frac{a}{b^2}x_1^2\right\}\mathbf{e}_2)\, ds$$

$$+ \frac{c_3}{Length}\int_{x_1=-b}^{x_2=b} \Phi_e\left(\mathbf{x}_e + x_1\mathbf{e}_1 - \left\{c - \frac{c}{b^2}x_1^2\right\}\mathbf{e}_2\right) ds \qquad (13)$$

where s corresponds to the arc length of the curve and *Length* to its total length. Note that scale independence is achieved by dividing line integrals by their total length and double integrals (over regions) by their area.

The minimization is done by steepest descent of the energy function in parameter space. It is assumed that preprocessing, or interactions between different templates (see section two), will allow the eye-template to start relatively near the correct position. Alternatively the eye template might locate several promising positions, probably at valleys in the intensity, and investigate them.

Thus the update rule for a parameter, for example r, is given by

$$\frac{dr}{dt} = -\frac{\partial E_C}{\partial r} \qquad (14)$$

These terms are explicitly calculated in Yuille et al. (1988).

Simulation Results for Eyes

The theory was tested on real images using a SUN4 computer. The valleys, peaks, and edges are first extracted and smoothed (see Fig. 4). The template is then given initial parameter values, positioned in the image and allowed to deform itself using the update equations.

Some initial experimentation was needed to find good values for the coefficients and a number of problems arose. For example, the intensity and valley terms over the circle attempt to find the maximum value of the potential terms *averaged* inside the circle. This led to the circle shrinking to a point at the darkest part of the iris. This effect was countered by strengthening the edge terms, which pull the circle out to the edge between the iris and the whites of the eye. Another problem arose because the iris might also be partially hidden by the boundary of the eye, thus the part of the circle outside the boundary cannot be allowed to interact with the image. This can be dealt with by considering only the area of the circle inside the bounding parabolas.

The system worked well after good values were found for the coefficients. The templates usually converged to the eye provided they were started at or below it. The valleys from the eyebrows caused problems if the template was started above the eye.

The values of the coefficients changed automatically during the course of the program to define six distinct epochs:

1. The coefficients of the valley forces are strong and the force is calculated using the representation fields. The coefficients of the peak and edge forces are zero.

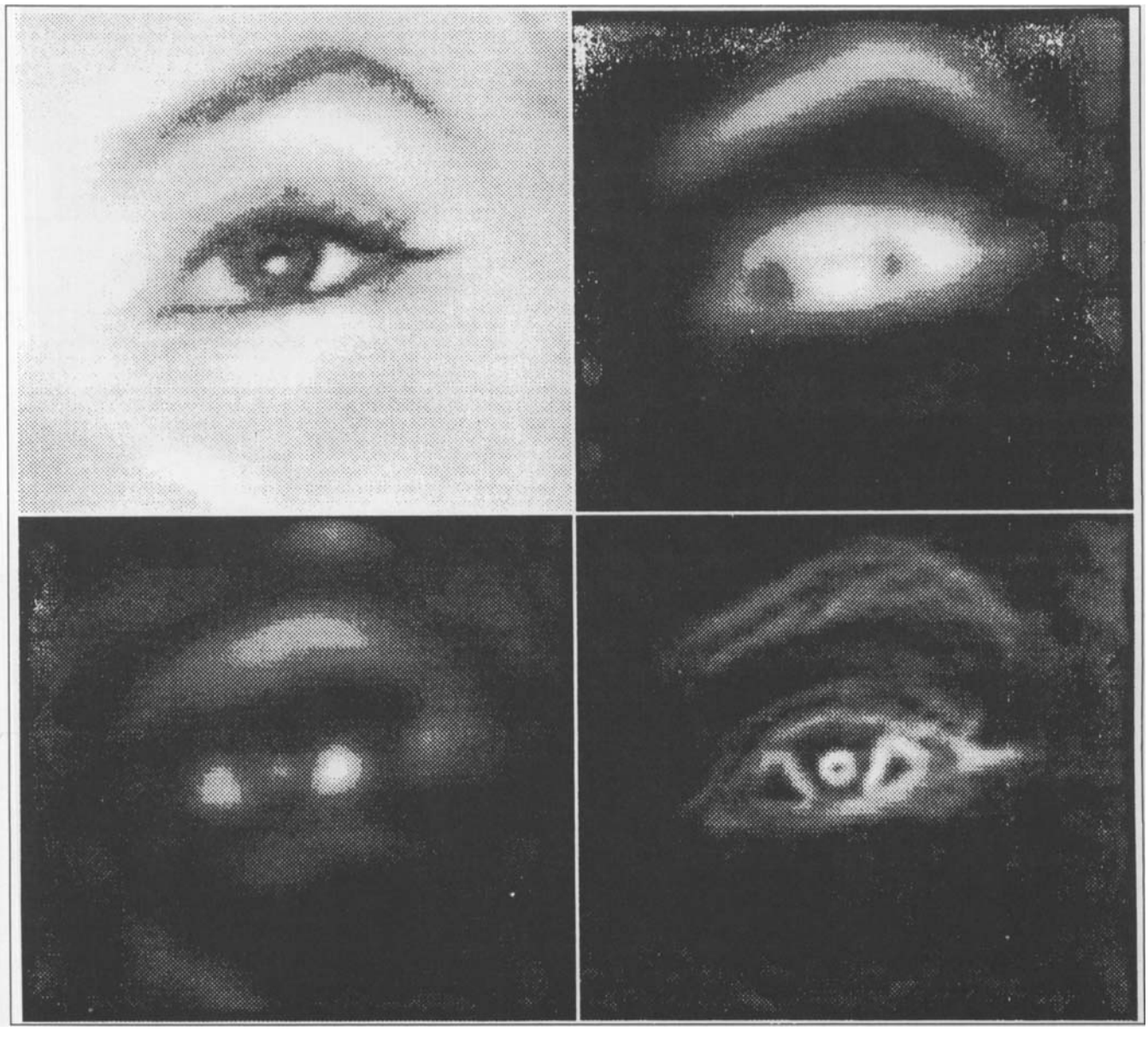

Figure 4. The valley, edge, and peak representations for the eye region.

During this epoch the valley forces pull the template to the eye.

2. The valley forces are calculated using the intensity field, with the peak and valley fields still zero. This helps scale the circle to the correct size of the iris.

3. The edge coefficients for the boundary of the circle increase. This fine tunes the size of the circle as it locks onto the iris.

4. The peak coefficients increase. This enables the peak forces, using the representation field, to rotate the template and get the correct orientation.

5. The peak forces are calculated with the intensity field. This helps adjust the size of the outer boundary of the template.

6. The coefficients of the edges of the boundary are increased. This fine tunes the positions of the boundaries.

The program changes epoch automatically when it has reached a steady state of the energy function with the appropriate coefficient values (i.e., when it thinks it has accomplished its goals for that epoch).

Figure 5 illustrates the program running in the different epochs. Note that the template can start some distance away from the eye, can scale the iris, rotate the eye, and lock onto the edges. The runtime for the program is between 5 and 10 min on a SUN4.

Note that in the above scenario the representation fields are used for long range attraction and the intensity fields do the fine scale tuning. If the initial guess for the

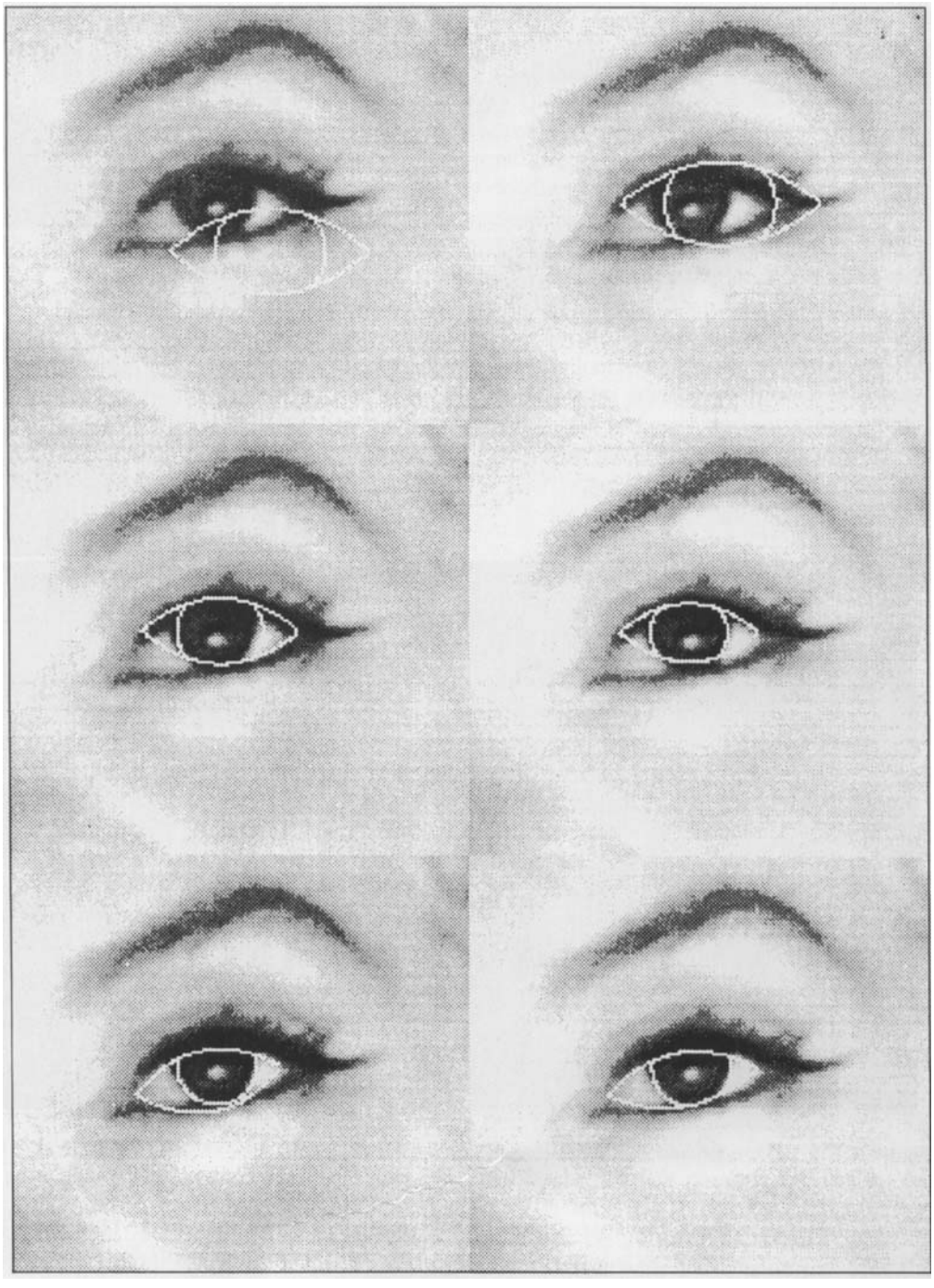

Figure 5. A dynamic sequence for the eye left to right and top to bottom. The first frame shows the initial configuration and the remaining frames show the results at the ends of the epochs.

location and size is accurate then the representation fields are unneccessary.

Summary of Feature Templates

Yuille et al. (1988) describes how this work can be extended to detect mouths. We define a parameterized template for the mouth and allow it to adjust itself to the image (see Fig. 6).

It seems relatively straightforward to find templates for the other "internal" features of the face, such as eyebrows, noses, chins, and moustaches. There has also been some promising initial work on extracting foreheads (P. Maragos, private communication). It is less clear how to generalize this idea to find "external" features such as the ears or hair. Possibly features of this type are best found at a coarser level.

The approach can be directly adapted to many other recognition problems. For example, Lipson, Yuille, O'Keefe, Cavanaugh, Taaffe, and Rosenthal (1990) describe a successful system for extracting trabecular bones from medical images.

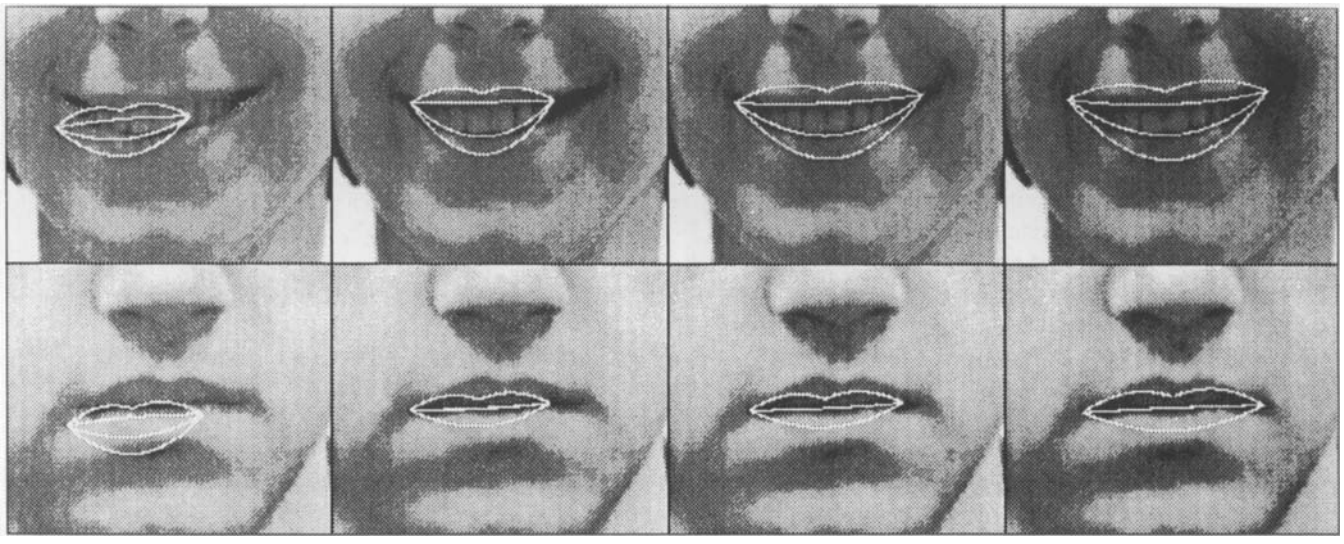

Figure 6. A dynamic sequence for the mouth template on an open and a closed mouth. In the upper picture the template is pulled in by the peak and edge forces from the teeth. In the lower picture the template is mainly pulled in by the valley forces and the region for the teeth vanishes.

In the experiments described above the initial values for the templates were selected by hand and were often chosen to make the task hard. For an automated system there are two extreme strategies. The first approach would be to run global face templates (see the second section) to determine likely positions for the features, which could then be checked using deformable feature templates. The second strategy, which is purely local, would be to use the representation fields to determine possible likely candidates. For example, the eye template might start at places where the valley representation was strong, thereby eliminating the need for the first epoch and enabling the template to act directly on the image intensity. Since there will probably be several candidates we would need to start several deformable templates off in parallel and see which gives the best results. This would require some criteria for selecting the best fit. A natural choice would be the one with the lowest final energy function. This, however, might need to be supplemented by taking into account the spatial relationships to other features and the a priori probability of the final parameter values. In some special cases it may be possible for the energy to be low but for the parameter values to be extremely unlikely. Such a situation can occur if the mouth template gets started on the eye and becomes grotesquely deformed (Yuille et al., 1988) (see Fig. 7).

Deformable feature templates give a promising approach to extracting features. A problem with the approach is the difficulty of extracting the representation fields automatically since the fields, in particular the peak fields, are sensitive to the scale of the morphological operators. This does not matter for the final measure of fit, which is based almost entirely on the intensity fields, but it would affect the matching strategy (the epochs). In the next section we describe a reformulation of deformable templates that makes them more robust and reliable.

ROBUST FEATURE TEMPLATES

The templates described in the previous section were designed on a somewhat ad hoc basis. There was a measure of fit, the energy function, but no explicit imaging model. There are also several situations for which the templates may fail, for example, if the mouth is smoking a cigarette. In this section we discuss current research, which attempts to put the templates in a more theoretical setting and make them more robust (Hallinan & Mumford, 1990). We will concentrate chiefly on the imaging model for the template and its measure of fit, with less emphasis on how the matching is done to minimize this measure.

This work draws on ideas developed in *Robust Statistics* (Huber, 1981; Rousseeuw, 1987; for other applications to vision see Pavlidis, 1986; McKendall & Mintz, 1989). To motivate these ideas consider the problem of estimating the mean from a set of samples $\{x_i\}$, $i = 1, \ldots, N$. The sample mean is defined to be

$$\bar{x} = (1/N) \sum_{i=1}^{N} x_i \tag{15}$$

and minimizes the least squares error

$$E(x) = \sum_{i=1}^{N} (x - x_i)^2 \tag{16}$$

The sample mean, however, is extremely sensitive to outliers. By introducing a new point x_{N+1} a sufficiently

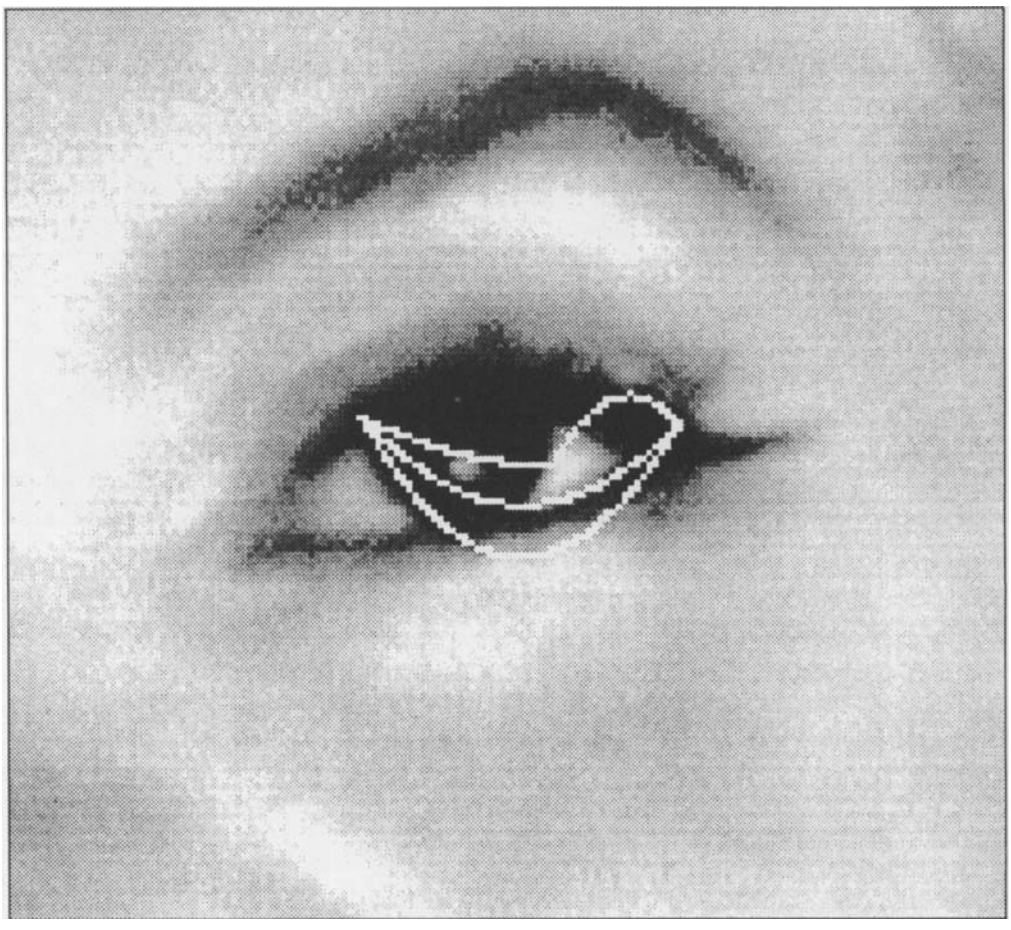

Figure 7. A grotesquely deformed template. See text.

long way away from the other points we can alter the value of the sample mean by an arbitrarily large amount. A robust technique for estimating the mean should be relatively independent of such outliers and should also enable us to identify the outliers themselves.

There are several possible robust techniques. Perhaps the most simple is least trimmed squares. For each value of x we order the *residuals* $r_i = (x - x_i)^2$ so that $r_{(1)} \leq r_{(2)} \leq \ldots$. We now minimize the least trimmed squares (LTS)

$$E_{\mathrm{LTS}}(x) = \sum_{i=1}^{M} r_{(i)} \tag{17}$$

where M is an integer less than N giving the proportion of points that we wish to match. If $M = \alpha N$ then this is the least α-trimmed squares estimator, closely related to the α-trimmed mean estimator. M may be altered adaptively.

Thus minimizing $E_{\mathrm{LTS}}(x)$ with respect to x gives us the estimate of the mean for the best M points. We can simultaneously get an estimate of the variance of the residuals of these points. The remaining $N - M$ points are treated as outliers if their residuals are significantly larger than the variance.

This example demonstrates two important aspects of *Robust Statistics*: (1) finding an estimate of a quantity by discarding, or reducing the influence of, some of the data, and (2) analyzing the residuals of the discarded data.

Hallinan and Mumford propose adapting these techniques to produce a Robust Deformable Template. To illustrate these ideas we consider their reformulation of the eye template.

Their model of the ideal eye has essentially the same geometry as the old eye template (see the third section). In their ideal imaging model the iris and the whites of the eyes are both assumed to have constant image intensity. This can be distorted by the addition of noise and by the overlay of occluding objects (see Fig. 8). Note that for blue eyes the iris and pupil have different intensity and a more complex model is needed. The use of robust matching criteria means that the system is not very sensitive to the precise form of the imaging model.

The measures of fit aim to find the parameters of the template that minimize the mean in the iris region, maximize the mean for the whites of the eyes, and maximize the mean edge strength at the boundaries. This is somewhat similar to the measure used by the eye template in the third section for the fine scale matching when the representation fields are not used. The difference lies in the use of an underlying imaging model, robust methods for calculating the means, residual analysis, and the use

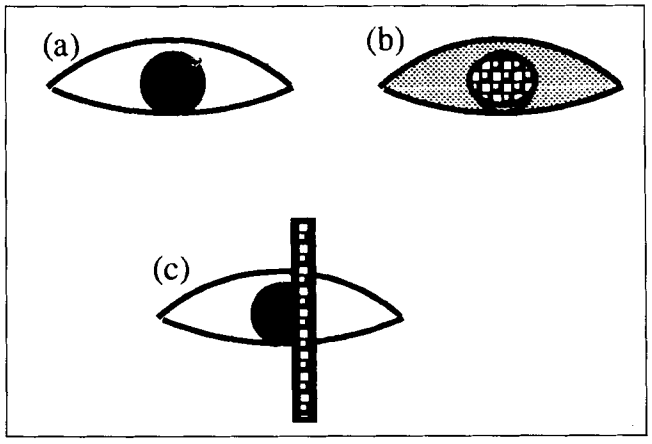

Figure 8. **(a)** The ideal eye has constant low intensity in the iris and constant high intensity in the whites. **(b)** The ideal eye with noise superimposed. **(c)** The ideal eye with an occluding bar.

of the variance to determine the degree of flatness of the regions. The goodness of fit can be defined in a dimensionless, parameter-free way as the percentage of variance accounted for by the imaging model in a fixed proportion $(1 - \alpha)$ of the pixels.

The use of robust techniques, in particular the α-trimmed mean, enable the means to be estimated correctly and the template matched even when a significant part of the eye is occluded (see Fig. 8c). Intuitively the measures reward good partial matches more than they penalize poor partial matches. Hallinan and Mumford also suggest that residual analysis could be applied so that matches could be strengthened if the discarded data could be interpreted as an occluding object. Thus the image intensity within the portion of the bar in Figure 8c, which overlaps with the eye might be identified as outliers. Further analysis could show that these outliers form a coherent structure, the bar, occluding the eye.

The chief advantage of this approach is that it specifies a precise imaging model, allowing for gross distortions such as the occluding bar, and designs the fitness measure accordingly. It is hence possible to say precisely for which class of stimuli the method will work. Of course it is necessary to do experiments to test how accurate the imaging models are.

CONCLUSION

Deformable templates offer a promising and conceptually attractive way for locating features, and sets of features with given spatial relations, by exploiting prior knowledge. The parameters of the templates can then be used to describe and recognize faces.

The methods described here are consistent with two extreme strategies for face description. In the bottom up strategy individual features are located before their spatial relations are determined. In the top down approach the spatial relations are used to guide the location of individual features.

The templates we have described are essentially "selfish." They try to grab the parts of the image that they can explain and ignore the rest. One can contrast this with the more Bayesian approach used by Chow et al. (1989), which attempts to extract the outline of a hand by assuming that the hand and the background have constant, though unknown, image intensities with superimposed gaussian noise. This approach attempts to explain the whole image and requires knowledge of the background as well as knowledge of the object being detected. In cases where knowledge of the background is unavailable the selfish template approach may be preferable.

Selfish templates alone, however, might not be sufficient for giving a good image interpretation. Instead they might need to be supplemented by "altruistic" templates that attempt to explain the whole image. Although selfish templates try to match from the template to the image the altruistic templates try to match from the image to the templates. Thus some of the altruistic templates may be only partially matched. The use of residual analysis might provide a link between these two template types. The altruistic templates might try to explain the parts of the image unexplained by the selfish templates.

The measures based on *Robust Statistics* enable the template to be matched even when part of it is occluded. Residual analysis may be able to isolate the occluder and describe it separately. By refining the imaging models and the robust measures of fit we expect to be able to reliably extract features from images of faces under a large variety of lighting conditions.

Deformable templates were invented as a technique for doing computer vision and were not primarily motivated by considerations from psychological and neuroscientific experiments. Nevertheless there are some intriguing connections to such experiments, which we are currently investigating. The deformable templates themselves have some similarities to mental images (Kosslyn, 1980). In a slightly different vein they would suggest an initial representation of a face somewhat similar to a caricature though with more information. This representation, the description generated by the local and global deformable templates, would include peaks and valleys, in addition to edges, and the spatial relations between features. It is interesting that, though it is usually harder to recognize someone from a caricature than from a photograph, using caricatures can speed up reaction time for very familiar faces (Rhodes, Brennan, & Carey, 1987). The global face models might be related to the results of Haig (1984) on the sensitivity of observers to the spatial relations between features.

Acknowledgments

I would like to thank the National Science Foundation for Grant IRI-9002141 and the Brown, Harvard, and M.I.T. Center for Intelligent Control Systems for United States Army Research Office Grant DAAL03-86-C-0171. Conversations with Peter Hal-

linan, David Mumford, Mark Nitzberg, Taka Shiota, and Udo Wehmeier were extremely useful.

Reprint requests should be sent to Alan Yuille, Division of Applied Sciences, G 12E Pierce Hall, Harvard University, Cambridge, MA 02138.

REFERENCES

Bridgman, G. B. (1973). *Constructive anatomy.* New York: Dover.

Burr, D. J. (1981a). A dynamic model for image registration. *Computer Graphics and Image Processing, 15,* 102–112.

Burr, D. J. (1981b). Elastic matching of line drawings. *IEEE Transactions Pattern Analysis and Machine Intelligence, PAMI-3* (6), 708–713.

Burt, P., & Adelson, E. (1983). The Laplacian as a compact image code. *IEEE Transactions on Communications, 31,* 532–540.

Chow, Y., Grenander, U., & Keenan, D. M. (1989). Hands: A pattern theoretic study of biological shape. Monograph. Brown University, Providence, RI.

Durbin, R., Szeliski, R., & Yuille, A. L. (1989). An analysis of the elastic net approach to the travelling salesman problem. *Neural Computation, 1,* 348–358.

Durbin, R., & Willshaw, D. J. (1987). An analogue approach to the travelling salesman problem using an elastic net method. *Nature (London) 326,* 689–691.

Fischler, M. A., & Elschlager, R. A. (1973). The representation and matching of pictorial structures. *IEEE Transactions on Computers, 22,* 1.

Goldstein, A. J., Harmon, L. D., & Lesk, A. B. (1972). Man-machine interaction in human-face identification. *The Bell System Technical Journal, 51* (2), 399–427.

Grenander, U. (1989). The 1985 Rietz Lecture: Advances in pattern theory. *The Annals of Statistics, 17* (1), 1–30.

Haig, N. D. (1984). The effects of feature displacement on face recognition. *Perception, 13,* 505–512.

Hallinan, P. W., & Mumford, D. (1990.) In preparation.

Huber, P. J. (1981). *Robust statistics.* New York: John Wiley.

Kanade, T. (1977). *Computer recognition of human faces.* Basel and Stuttgart: Birkhauser Verlag.

Kass, M., Witkin, A., & Terzopoulos, D. (1987). Snakes: Active contour models. *Proceedings of the First International Conference on Computer Vision,* London.

Knoerr, A. (1989). Global models of natural boundaries: Theory and applications. Pattern Analysis Tech. Rep. No. 148. Brown University, Providence, RI.

Kosslyn, S. (1980). *Images and mind.* Cambridge, MA: Harvard University Press.

Lipson, P., Yuille, A. L., O'Keefe, D., Cavanaugh, J., Taaffe, J., & Rosenthal, D. (1990). Deformable templates for feature extraction from medical images. *Proceedings of the First European Conference on Computer Vision,* Antibes, France.

Maragos, P. (1987). Tutorial on advances in morphological image processing and analysis. *Optical Engineering, 26,* 623–632.

McKendall, R., & Mintz, M. (1989). Robust fusion of location information. Preprint. Department of Computer and Information Science, University of Pennsylvania.

Pavlidis, T. (1986). A critical survey of image analysis methods. *Proceedings of the Eighth International Conference on Pattern Recognition,* Paris.

Rhodes, G., Brennan, S., & Carey, S. (1987). Identifying and rating: Implications for mental representations of faces. Preprint.

Rousseeuw, P. J. (1987). *Robust regression and outlier detection.* New York: John Wiley.

Serra, J. (1982). *Image analysis and mathematical morphology.* New York: Academic Press.

Terzopoulos, D., Witkin, A., & Kass, M. (1987). Symmetry-seeking models for 3D Object Recognition. *Proceedings of the First International Conference on Computer Vision,* London, June.

Turk, M., & Pentland, A. (1989). Face processing: Models for recognition. SPIE. Vol 1192. Intelligent robots and computer vision VIII: Algorithms and techniques.

Yuille, A. L., Cohen, D. S., & Hallinan, P. W. (1988). Facial feature extraction by deformable templates. Harvard Robotics Lab. Tech. Rep. 88-2.