Methods of Global Optimization in the Tracking of Contours

Daniel Freedman and Michael S. Brandstein
Harvard University
Division of Engineering and Applied Sciences
Cambridge, MA, USA 02138
{freedman, msb}@hrl.harvard.edu

Abstract

A new method for tracking contours of moving objects in clutter is presented. For a given object, a model of its contours is learned from training contours in the form of a subset of curve space. Complexity is added to the contour model by analyzing rigid and non-rigid transformations of contours separately. In the course of tracking, a very large number of potential curves are typically observed due to the presence of extraneous edges in the form of clutter; the learned model guides the algorithm in picking out the correct one. The algorithm is posed as a solution to a minimization problem; theoretical results on how to achieve the global minimum to within a certain resolution, and the complexity of this operation, are presented. Experimental results applying the proposed algorithm to the tracking of a flexing finger and to a conversing individual's lips are also presented.

1 Introduction

1.1 Review of Existing Approaches

The goal in contour tracking is to follow the silhouette of an object as it moves through a video-stream. To do so, the deformable template approach [6] minimizes, for each frame, an energy function which is specific to the geometry of the tracked object. Elastic snakes [5], by contrast, minimize a more general energy function, which has terms representing elastic and tensile energy to ensure that the snake is smooth, and an image-dependent term that pushes the snake towards the feature of interest. The Kalman tracker [1] requires a learned linear stochastic dynamical model which describes the evolution of the contour to be tracked. Assuming that the observation of the contour has been corrupted by Gaussian noise, the conditional density of the contour given all past observations may be found, and then used to estimate the contour position. The condensation tracker [2] also assumes a dynamical model describing contour motion is known and that imprecise observations are made. However, both the dynamical system and the observation process may be completely general and the conditional density may be propagated forward in time using the numerical technique known as the "condensation" method. This density may then be used for estimating the current contour.

1.2 The Problem

In either tracking an object through a video-stream, or localizing it within a single image, the approach taken will be to focus entirely on the object's contour, or outline. Thus, the problem of tracking or localizing reduces to one of finding the "correct" curve in the image, i.e. the curve which corresponds to the object of interest. Suppose it is possible to generate two sets of curves. One set, $\tilde{E}$, represents all of the curves that can be generated from the connection of edge-points in the image; a description of how to produce such a set follows in section 1.3. The other set, $\tilde{C}$, represents all of the curves that correspond to the particular geometry of the object being tracked or localized; that is, this set $\tilde{C}$ contains all of the information about the shape of the object's silhouette. A discussion of this set is provided at the end of section 1.3. Given these two sets, a sensible problem to solve is

$$\min_{\tilde{e}\in\tilde{E},\tilde{c}\in\tilde{C}} ||\tilde{e}-\tilde{c}||$$

where $||\cdot||$ is the L_2 norm. The idea is straightforward: the minimization over the two arguments ensures that the "observed curve" $\tilde{e}$ chosen from all of the possible curves in the image best matches the model of the object being located or tracked, as embodied in the set $\tilde{C}$. The tracked or localized object is taken to be $\tilde{e}^*$, the minimizing argument. An earlier attempt, on the part of the authors, to solve a similar problem is contained in [4]; the approach presented in the following sections, however, is much more flexible and robust.

1.3 The Sets $\tilde{E}$ and $\tilde{C}$

Focus first on tracking; in this case, the set $\tilde{E}$, of curves constructed from edge-points in the image, is generated as follows. At N equally spaced points along the detected contour of the previous frame, edge-search takes place in circular regions (in the image of the *current* frame). In each of these regions, a number of edge-points are detected; denote the set of edge-points detected in the n^{th} region by E_n. An element $\tilde{e} \in \tilde{E}$ may be constructed as follows:

- take one edge-point $e_n \in E_n$ from each region $n = 1, \ldots, N$;
- smoothly interpolate the set of edge-points $e_1, \ldots, e_N$ into a curve $\tilde{e}$.

(The method of interpolation will not concern us here.) Thus, the set $\tilde{E}$ is in one-to-one correspondence with the set $E \equiv E_1 \times \cdots \times E_N$. Suppose there are M edge-points detected per site, i.e. $|E_n| = M \; \forall n$ (in reality, of course, $|E_{n_1}| \neq |E_{n_2}|$); then the size of the set of observed curves is $|\tilde{E}| = |E| = M^N$.

In the case of object localization, the same principle may be applied for finding observed curves, only now there are no natural candidates for search regions, as one cannot initiate search based on localization in the previous frame (since there *is* no previous frame). Instead, let the set of *all* edge-points in the image be denoted Υ; then take $E_n = \Upsilon \; \forall n$, and E and $\tilde{E}$ are constructed from the $\{E_n\}_{n=1}^N$ as before. That is, an "observed curve" may be generated by selecting a subset N out of any of the edge-points in the image, in a particular order, and then interpolating between them. (For simplicity, repeated points are allowed.) Again, $|\tilde{E}| = |E| = M^N$, but $M = |\Upsilon|$ in this case is much larger than M is for tracking.

The set $\tilde{C}$ is generated from training curves before the algorithm is run. It is assumed that this set of learned curves is a finite dimensional manifold (as it will be in all cases that will be practically encountered), has dimension σ, and may be specified parametrically as

$$\tilde{C} = \{\tilde{c}(u) : u \in U\}$$

where U is some known, σ-dimensional, real, compact, convex set (for example, $U = [0,1]^\sigma$), and $\tilde{c}(\cdot)$ is a function which maps the points in U to points in curve space. Some of the parameters may represent familiar transformations; for example, $u_1, \ldots, u_6$ could represent a subset of the affine transformations. In fact, given a particular group of transformations, a useful learning scheme would involve finding an invariant to this group for each of the training curves, and learning on these invariants; then, the group of transformations can be added in as suggested above. In this case, learning is only used to understand the non-affine deformations, for which there is no guiding theoretical structure. For the moment, no particular learning method is outlined; instead, it is simply assumed that $\tilde{C}$ as specified above is known.

1.4 Recasting the Problem

Using the parametric form for $\tilde{C}$ allows the problem to be re-written

$$\min_{\tilde{e} \in \tilde{E}, u \in U} ||\tilde{e} - \tilde{c}(u)||$$

However, approximating the square of the L_2 norm by its Riemann sum gives

$$\begin{aligned} ||\tilde{e} - \tilde{c}||^2 &= \int_0^L ||\tilde{e}(s) - \tilde{c}(s)||^2 ds \\ &\approx \frac{L}{N} \sum_{n=1}^N ||e_n - c_n||^2 \end{aligned}$$

where $e_n = \tilde{e}(s_n), c_n = \tilde{c}(s_n)$, and $s_n = \frac{L(n-1)}{N-1}$. Note that $e_1, \ldots, e_N$ is simply the set of edge-points, culled from the sets $E_1, \ldots, E_N$, which were interpolated to give $\tilde{e}$; sampling $\tilde{e}$ gives back the original points. Denoting $e = (e_1, \ldots, e_N) \in \Re^{2N}$ and similarly for c, then the minimization problem may be approximated well by

$$\min_{e \in E, u \in U} ||e - c(u)||$$

if N is sufficiently large. Note that the norm in the above is the now the normal Euclidean norm in $\Re^{2N}$, $E = E_1 \times \cdots \times E_N$ as before, and $c(\cdot) : \Re^\sigma \to \Re^{2N}$.

The recast problem is still not obviously amenable to solution, as E is still discrete and very large, while U is continuous. Below, an algorithm will be proposed for solving for the global optimum. In particular, if d^* is the value of the global minimum, then the algorithm will be shown to give a value of at most $d^* + \Delta d$, for a specified Δd. Further, complexity bounds on the algorithm, in terms of M, N, and Δd will be established. The essence of the algorithm is contained in the following theorem.

2 A Theorem on Global Optimization

Before stating the theorem, it will be necessary to make two definitions.

Definition: V is said to be an ε-**cover** of the compact set U if $\forall u \in U$, $\exists v \in V$ such that $||v - u|| \leq \varepsilon$, and ε is the smallest such value. Alternatively, $\varepsilon = \max_{u \in U} [\min_{v \in V} ||v - u||]$. (Note that the maximum is well-defined since U is compact.)

Definition: Let $H(u) = \frac{\partial c}{\partial u}$, so that $H(u) \in \Re^{2N \times \sigma}$. Let $\lambda_1(u)$ be the largest eigenvalue of the $\sigma \times \sigma$ matrix $H^T(u)H(u)$. Then for any $Y \subset U$, define $A(Y) = [\max_{u \in Y} \lambda_1(u)]^{1/2}$.

We are now ready to state the theorem which will allow us to attack the tracking / localization problem, $\min_{e \in E, u \in U} ||e - c(u)||$. The import of the theorem will be discussed after its formal statement.

Theorem: Let V be any ε-cover of U. Further, let $d^* = \min_{e \in E, u \in U} ||e - c(u)||$ and let $d^\dagger = \min_{u \in U} ||e^\dagger - c(u)||$, where $e^\dagger = \arg\min_{e \in E} (\min_{v \in V} ||e - c(v)||)$. If $\Delta d = d^\dagger - d^*$, then

$$0 \leq \Delta d \leq \frac{3A^2(U)\varepsilon^2}{d^*} + 2A(U)\varepsilon.$$

This theorem presents a problem whose solution is amenable, and compares the objective function value it gives compared to the optimal, d^*. In particular, the problem $\min_{e \in E, v \in V} ||e - c(v)||$ can be solved, albeit inefficiently, by exhaustive search through the two discrete sets E and V. If the e-minimizing argument is labelled $e^\dagger$, then the quantity $d^\dagger = \min_{u \in U} ||e^\dagger - c(u)||$ is of interest; the fact that the minimizing u is never solved for does not matter, since our contour estimate is based on $e^\dagger$ rather than $c(u^\dagger)$ (see section 1.2). The theorem gives an upper bound on how far away $d^\dagger$ can be from d^*; this bound depends critically on ε, a parameter which indicates how finely V samples U.

3 Proof of the Theorem

Begin by considering only two sampled observed curves, e_1 and e_2. Make the following definitions: for $i = 1, 2$, let

- $u_i = \arg\min_{u \in U} ||e_i - c(u)||, \quad d_i = ||e_i - c(u_i)||$
- $\tilde{u}_i = \arg\min_{v \in V} ||e_i - c(v)||, \quad \tilde{d}_i = ||e_i - c(\tilde{u}_i)||$
- $\hat{u}_i = \arg\min_{v \in V} ||u_i - v||, \quad \hat{d}_i = ||e_i - c(\hat{u}_i)||$

Then:

$$\begin{aligned} d_2^2 - d_1^2 &= ||e_2 - c(u_2)||^2 - ||e_1 - c(u_1)||^2 \\ &\leq ||e_2 - c(\hat{u}_2)||^2 + ||c(\hat{u}_2) - c(u_2)||^2 \\ &\quad - ||e_1 - c(u_1)||^2 \\ &\leq ||e_2 - c(\hat{u}_2)||^2 + ||c(\hat{u}_2) - c(u_2)||^2 \\ &\quad - ||e_1 - c(\hat{u}_1)||^2 + ||c(\hat{u}_1) - c(u_1)||^2 \\ &= \hat{d}_2^2 - \hat{d}_1^2 + ||c(\hat{u}_1) - c(u_1)||^2 + ||c(\hat{u}_2) - c(u_2)||^2 \end{aligned}$$

where the second and third inequalities are both applications of the triangle inequality. Now:

1. $\hat{d}_1 \geq \tilde{d}_1$ by definition, so $-\hat{d}_1^2 \leq -\tilde{d}_1^2$

2. Expanding $\hat{d}_2^2 = ||e_2 - c(\hat{u}_2)||^2$ gives

   $$\hat{d}_2^2 = ||e_2 - [c(u_2) + H(\breve{u}_2)(\hat{u}_2 - u_2)]||^2$$

   where $H(u) = \frac{\partial c}{\partial u}$ and $\breve{u}_2 \in U$. This is the multivariable mean value theorem, which is valid due to the convexity of U [3]. Thus,

   $$\begin{aligned} \hat{d}_2^2 = &||e_2 - c(u_2)||^2 + ||H(\breve{u}_2)(\hat{u}_2 - u_2)||^2 \\ &+ 2(e_2 - c(u_2))^T H(\breve{u}_2)(\hat{u}_2 - u_2) \end{aligned}$$

   (a) $||e_2 - c(u_2)||^2 = d_2^2 \leq \tilde{d}_2^2$

   (b) $||H(\breve{u}_2)(\hat{u}_2 - u_2)||^2 = (\hat{u}_2 - u_2)^T H^T(\breve{u}_2) H(\breve{u}_2)$ $(\hat{u}_2 - u_2)$. Since V is an ε-cover of U, $\exists v$ such that $||u_2 - v|| \leq \varepsilon$. But by definition $\hat{u}_2 = \arg\min_{v \in V} ||u_2 - v||$; thus, $||\hat{u}_2 - u_2|| \leq \varepsilon$. But then

   $$\begin{aligned} (\hat{u}_2 - u_2)^T H^T(\breve{u}_2) H(\breve{u}_2)(\hat{u}_2 - u_2) &\leq \lambda_1(\breve{u}_2)\varepsilon^2 \\ &\leq \left(\max_{u \in U} \lambda_1(u)\right)\varepsilon^2 \equiv A^2(U)\varepsilon^2 \end{aligned}$$

   where $\lambda_1(\cdot)$ and $A(\cdot)$ are defined as before.

   (c) Finally,

   $$\begin{aligned} &(e_2 - c(u_2))^T H(\breve{u}_2)(\hat{u}_2 - u_2) \\ &\quad \leq |(e_2 - c(u_2))^T H(\breve{u}_2)(\hat{u}_2 - u_2)| \\ &\quad \leq ||e_2 - c(u_2)|| ||H(\breve{u}_2)(\hat{u}_2 - u_2)|| \end{aligned}$$

   where the latter inequality is due to the Cauchy-Scwartz inequality. But $||e_2 - c(u_2)|| = d_2$ and from the previous argument $||H(\breve{u}_2)(\hat{u}_2 - u_2)|| \leq A(U)\varepsilon$. Thus, $(e_2 - c(u_2))^T H(\breve{u}_2)(\hat{u}_2 - u_2) \leq d_2 A(U)\varepsilon$.

3. Using the mean value theorem once again,

   $$\begin{aligned} ||c(\hat{u}_i) - c(u_i)||^2 &= ||c(u_i) + H(\breve{u}_i)(\hat{u}_i - u_i) - c(u_i)||^2 \\ &= ||H(\breve{u}_i)(\hat{u}_i - u_i)||^2 \\ &\leq A^2(U)\varepsilon^2 \end{aligned}$$

Thus,

$$\begin{aligned} d_2^2 - d_1^2 &\leq \tilde{d}_2^2 + A^2(U)\varepsilon^2 + 2d_2A(U)\varepsilon - \tilde{d}_1^2 + A^2(U)\varepsilon^2 \\ &+ A^2(U)\varepsilon^2 = \tilde{d}_2^2 - \tilde{d}_1^2 + 3A^2(U)\varepsilon^2 + 2d_2A(U)\varepsilon \end{aligned}$$

The above inequality is valid for any e_1 and e_2. Now, consider in particular $e_1 = e^* = \arg\min_{e \in E}[\min_{u \in U} ||e - c(u)||]$ and $e_2 = e^\dagger = \arg\min_{e \in E}[\min_{v \in V} ||e - c(v)||]$. Then by definition, $\tilde{d}_2 \leq \tilde{d}_1$, so that

$$d^{\dagger 2} - d^{*2} \leq 3A^2(U)\varepsilon^2 + 2d^\dagger A(U)\varepsilon$$

Experiment	N	D	Video Rate	Resolution	Running Sequence	Accuracy
Finger	80	20	30 Hz	320 by 240	202 frames = 6.7 s	100 %
Lips	80	20	13 Hz	320 by 240	130 frames = 10.0 s	94 %

Table 1. Summary of the experiments.

$$d^\dagger - d^* \leq \frac{3A^2(U)\varepsilon^2 + 2d^\dagger A(U)\varepsilon}{d^* + d^\dagger} = \frac{\frac{3A^2(U)\varepsilon^2}{d^\dagger} + 2A(U)\varepsilon}{1 + \frac{d^*}{d^\dagger}}$$

However, $1/(1+\frac{d^*}{d^\dagger}) \leq 1$ since $d^* \geq 0$, so that

$$\Delta d \leq \frac{3A^2(U)\varepsilon^2}{d^\dagger} + 2A(U)\varepsilon \leq \frac{3A^2(U)\varepsilon^2}{d^*} + 2A(U)\varepsilon \quad \blacksquare$$

4 Complexity and Implementation

The complexity of the optimization procedure is as follows. With no modification, the problem

$$\min_{e \in E, v \in V} ||e - c(v)||$$

has complexity $O(M^N|V|)$ since $|E| = M^N$. However, if the problem is solved as

$$\min_{v \in V} \left[\min_{e \in E} ||e - c(v)|| \right]$$

and it is noted that

$$\min_{e \in E} ||e - c(v)||^2 = \min_{e_1 \in E_1, \ldots, e_N \in E_N} \sum_{n=1}^{N} ||e_n - c_n(v)||^2$$
$$= \sum_{n=1}^{N} \min_{e_n \in E_n} ||e_n - c_n(v)||^2$$

then the complexity is reduced to $O(MN|V|)$ (since the latter step has a complexity of $O(MN)$). Further, using a result from computational geometry, it can be shown that each minimization of the form $\min_{e_n \in E_n} ||e_n - c_n(v)||$ can be performed with $O(\log M)$ complexity, leading to an overall complexity of $O(N|V| \log M)$. (Note: in order to gain this log factor, it is necessary to incur $O(M \log M)$ in overhead to calculate the relevant Voronoi diagram; however, this is negligible in the scheme of things.) It is useful to convert the complexity $O(N|V| \log M)$ into an expression which depends on $M, N,$ and Δd. Use a dimensional argument. Let V be an ε-covering of U; then using something akin to sphere-packing, it is clear that $vol(U) \approx |V|\varepsilon^\sigma$, where $\sigma = dim(U) = dim(C)$. That is, $|V| \propto \varepsilon^{-\sigma}$. Now, assuming that Δd is fairly small, then it can be shown that ε is fairly small, so that the upper bound on Δd from the optimization theorem is proportional to ε (that is, the term in ε^2 drops out). In this case, the algorithm has complexity $O(N \Delta d^{-\sigma} \log M)$.

5. Results and Conclusions

Two sets of results are presented to illustrate the effectiveness of the proposed tracker: a flexing finger and a speaker's lips. A summary is given in Table 1. In both cases, $\tilde{C}$ was learned in the following manner:

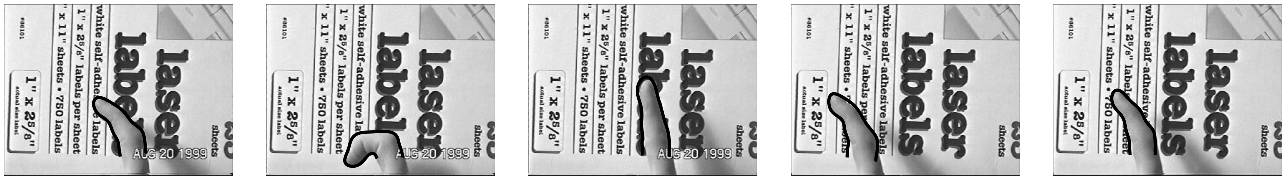
- Each training curve, represented as the $2D$ coefficients in a pair of D^{th} order Legendre polynomial expansions (one each for x and y), was transformed into its euclidean-similarity invariant, represented in the same basis.

- A one dimensional manifold was learned by smoothly interpolating through all of the invariants. This degree of freedom is captured in the variable u_1.

- 4 extra dimensions were then added, corresponding to the group of euclidean similarity transformations: translation in both x- and y-directions, rotations, and scaling. These degrees of freedom are represented by u_2, u_3, u_4, u_5.

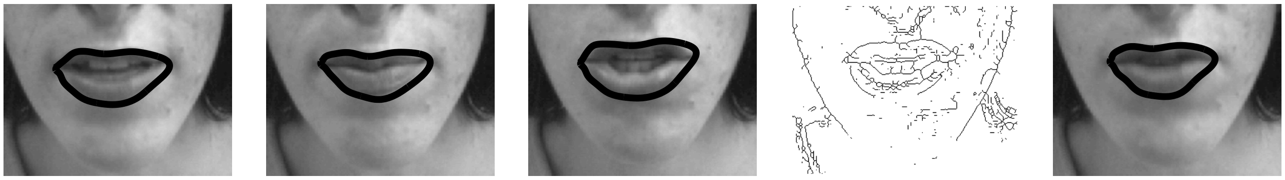
- Thus, both $\tilde{C}$ and C are five-dimensional manifolds. U is chosen to be $[0,1]^5$ for convenience.

The edge-map in the case of the finger was generated from the gray-scale intensity; clutter is in the form of both the background writing (much of which is small, and therefore leads to many extraneous edges) as well as the self-clutter of the doubled over finger. A sequence of tracked frames is shown in Figure 1; in this instance, the tracker got all 202 tracked frames close to correct. Note that motion consists of a combination of non-rigid deformations (flexing) as well as rigid motions (translation); the tracker is successful with both. In the case of the speaker's lips, the edge-map was generated from the green portion of the RGB image, which has slightly better contrast than the intensity; in addition, lipstick was used to help highlight contrast. Clutter is clearly visible in the edge-map shown in Figure 2, due to the detection of many extraneous edges, as well as the fact that over the search range the lips interfere with each other. A sequence of tracked frames is shown in Figure 2, and the tracker got 94% of the tracked frames correct; however, equally important as this high success rate is the ability to recover from the occasional error, as shown in Figure 3. Full video sequences of both tracking experiments can be viewed at http://himmel.hrl.harvard.edu/daniel/research.html.



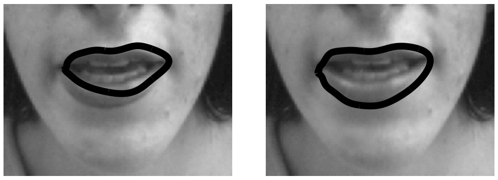
frame 30 frame 45 frame 80 frame 154 frame 201

Figure 1. Tracking a flexing and translating finger.



frame 17 frame 18 frame 75 edge-map for frame 75 frame 130

Figure 2. Tracking a speaker's lips.



frame 48 frame 49

Figure 3. Recovering from mistakes.

In the light of these successful experimental results, it is worth noting some of the advantages that are presented by this algorithm over other contour tracking approaches. As opposed to the deformable template approach, there is no need for hand-constructed models of the object's geometry; rather this is learned. Whereas elastic snakes use no special information about the object under study, the learned information used by the subset tracker allows for more accurate tracking. Furthermore, the subset tracker is computationally less burdensome than these two types of trackers. The subset tracker deals well with clutter, which is a main failing of the Kalman tracker. All of the advantages referred to above are matched by those of the condensation tracker. However, the condensation tracker relies on learned dynamical information, as opposed to the more basic shape information learned here. There are many situations in which the available training curves, which are used for learning prior to the running the algorithm, may be sufficient for learning the space in which object "lives," but are insufficient for learning the dynamics of the object.

Directions for future research include the development of an efficient multistage algorithm for maximizing over the set $V \subset U$; such an algorithm would have a tree-like, or coarse-to-fine structure, which allows subsets of U, in which the minimand cannot possibly reside, to be eliminated as the algorithm progresses. In addition, an algorithm for learning a multidimensional manifold would be of benefit. In the current experiments, one-dimensional manifolds was learned in invariant space; however, it is quite likely that the true manifolds were of higher dimension. (In this scenario, the one-dimensional manifold is simply a subset of the higher dimensional manifold.) Success in this area would also allow for more efficient implementation of the algorithm. Finally, the algorithm may be extended to the task of object localization, in which an object is to be located within a single image. Edge-search can no longer be initiated at the previous frame's contour estimate; thus, in principle any edge-point in the image may be potentially part of the relevant curve. The ability to search through the resulting huge space of observed curves relies on the $\log M$ term in the complexity, as discussed in section 4.

References

[1] A. Blake, R. Curwen, and A. Zisserman. A framework for spatio-temporal control in the tracking of visual contours. *Int. J. Comp. Vis.*, 11(2):127–145, 1993.

[2] A. Blake and M. Isard. Condensation - conditional density propagation for visual tracking. *Int. J. Comp. Vis.*, 29(1):5–28, 1998.

[3] W. Boothby. *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press, San Diego, 2nd edition, 1986.

[4] D. Freedman and M. Brandstein. A subset approach to contour tracking in clutter. In *Proceedings ICCV*, pages 242–247, 1999.

[5] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: active contour models. In *Proc. 1st Intern. Conf. Comput. Vis.*, London, June 1987.

[6] A. Yuille, P. Hallinan, and D. Cohen. Feature extraction from faces using deformable templates. *Int. J. Comp. Vis.*, 8(2):99–112, 1992.