

Short Papers

Object Tracking Using Deformable Templates

Yu Zhong, Anil K. Jain, *Fellow, IEEE*, and
M.-P. Dubuisson-Jolly, *Member, IEEE*

Abstract—We propose a novel method for object tracking using prototype-based deformable template models. To track an object in an image sequence, we use a criterion which combines two terms: the frame-to-frame deviations of the object shape and the fidelity of the modeled shape to the input image. The deformable template model utilizes the prior shape information which is extracted from the previous frames along with a systematic shape deformation scheme to model the object shape in a new frame. The following image information is used in the tracking process: 1) edge and gradient information: the object boundary consists of pixels with large image gradient, 2) region consistency: the same object region possesses consistent color and texture throughout the sequence, and 3) interframe motion: the boundary of a moving object is characterized by large interframe motion. The tracking proceeds by optimizing an objective function which combines both the shape deformation and the fidelity of the modeled shape to the current image (in terms of gradient, texture, and interframe motion). The inherent structure in the deformable template, together with region, motion, and image gradient cues, makes the proposed algorithm relatively insensitive to the adverse effects of weak image features and moderate amounts of occlusion.

Index Terms—Tracking, image sequence deformable template, shape, texture, motion.

1 INTRODUCTION

OBJECT tracking is a challenging and important problem in computer vision. Due to the nonrigid nature of objects in most of the tracking applications, deformable models are appealing in tracking tasks [1], [2], [4], [6], [10], [12], [13], [17], [18], [19], [22], [23], [24] because of their capability and flexibility. We can tune the prior structure information and define the manner in which the template interacts with images to obtain a deformable model that is suitable for a particular application. Deformable contours, such as snakes [13], [20], [19] have been applied to track rigid and nonrigid objects [1], [2], [4], [10], [13]. One advantage of the force-driven snake model is that it can easily incorporate the dynamics derived from time-varying images. Kass et al. [13] have used snakes to track facial features such as lips in an image sequence. The estimated motion parameters of these features are then used to explain facial expressions, etc. Multiple snakes were later used by Terzopoulos and Waters [21] to track more articulated facial features. Leymarie and Levine [16] have used the snake model to track cells in biological image sequences. DeCarlo and Metaxas [6] have proposed a deformable face model which includes both shape and motion parameters and have applied it to track human faces. Point distribution based active shape models [5], [15] were also proposed by Kervrann and Heitz [14] to track objects in long

image sequences, where a point distribution is used to characterize the structure and variations in the object shape.

Besides the object boundaries, attention has been focused recently on tracking objects using multiple image cues. One advantage of deformable models is its flexibility in combining multiple sources of information. Fua and Leclerc [8] have used a combination of stereo, shading, and smoothness in addition to surface shape deformations to reconstruct object surfaces. Appearance information including greyscale or color have been used in conjunction with deformable shape models [15], [22]. Basclé and Deriche [3] have combined deformable region models and deformable contours in a sequential way to track moving objects, where a correlation-based region matching method was used in the first stage to roughly locate the objects, and a gradient-based contour model was then used to refine the tracking result. Color information has also been used to track nonrigid objects in real-time applications [7], [9].

In this paper, we use the prototype-based deformable template model [11] to track objects in image sequences. In our approach, prior knowledge of an object shape is described by a hand-drawn prototype template which consists of the object's representative contour/edges. A deformed template is obtained by applying a parameterized deformation transform on the prototype. The shape variations in an object class are achieved by imposing a probabilistic distribution on the deformation parameters. The deformable shape template interacts with the input image via a potential field computed from the salient image features, for example, edges. The matching results are then evaluated using the objective function value which takes into account both the shape deviation from the prototype and the fidelity of the deformed template to the input data. The prototype-based deformable model is appealing for object tracking due to the following reasons: 1) The object of interest in the image sequence can vary from frame to frame due to a change in the view point, the motion of the object, or the nonrigid nature of the object. These shape variations can be captured by the deformable shape model, 2) although the object shape varies from frame to frame, the overall structure of the object is generally maintained. The deformable shape model can capture this overall structure by using an appropriate prototype, and 3) the motion or deformation between two successive frames is not significantly large so the converged configuration in the current frame can be used to provide a reasonable initialization for the next frame. To track objects in an image sequence, we propose a novel matching criterion which integrates image gradient, interframe motion, and region correspondence to track objects in the sequences.

2 PROTOTYPE-BASED DEFORMABLE MODEL

The deformable-template matching approach in [11], [23] provides a paradigm for object matching of arbitrary shapes. The prototype-based template combines both the global structure information and local image cues to derive an interpretation. It consists of three components: 1) a contour template which describes the prior knowledge about the object shape (a prototype), 2) a parameterized transformation which is applied to the prototype to deform it, and 3) a probabilistic distribution on the deformation parameters which controls the variation in the deformable template. This

- Y. Zhong is with the National Robotics Consortium, Carnegie Mellon University, Pittsburgh, PA 15201. E-mail: zhongyu@cs.cmu.edu.
- A.K. Jain is with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824. E-mail: jain@cse.msu.edu.
- M.-P. Dubuisson-jolly is with Siemens Corporate Research, Princeton, NJ 08540. E-mail: jolly@scr.siemens.com.

Manuscript received 11 Feb. 1998; revised 20 Mar. 2000; accepted 20 Mar. 2000.

Recommended for acceptance by S. Sarkar.

For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 107793.

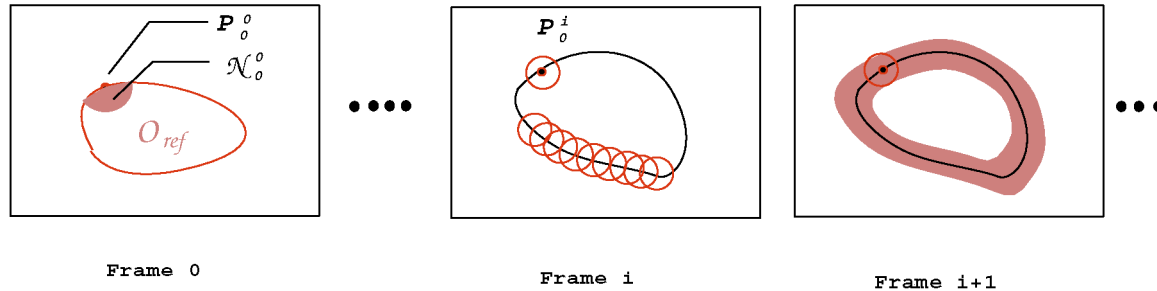


Fig. 1. Computing color (gray scale) distance. The detected object in the first frame is used as the reference object. For frame $i+1$, color (gray scale) distance is computed for the band (shaded region) around the detected object boundary in frame i .

model is able to deform the object shape to a certain degree to match the given image features.

The prototype template describes only one of the possible (though most likely) instances of the object shape. Therefore, it has to be deformed to match objects in images. The deformation transform determines the admissible deformation space, or equivalently, the possible shapes that a deformable template can take. In theory, the deformation transform can be any function which matches a 2D point to another 2D point, as it is used to approximate the displacement in a 2D plane. Let the deformation space be spanned by a discrete set of deformation basis: $\{\mathcal{F}_i(\mathbf{x}), i \in \mathcal{I}\}$, where \mathcal{I} is a index set for the basis functions. The displacement $\mathcal{D}(\mathbf{x})$ of a template point \mathbf{x} due to deformation parameters $\underline{\xi} = \{\xi_i, i \in \mathcal{I}\}$ is

$$\mathcal{D}_{\underline{\xi}}(\mathbf{x}) = \sum_{i \in \mathcal{I}} \xi_i \cdot \mathcal{F}_i(\mathbf{x}). \quad (1)$$

A good deformation transform should be capable of representing a variety of shape variations, providing a concise description, preferably with a computational advantage, and preserve the smoothness and connectivity of the template. We have used three different deformation transforms to span the displacement field, namely, the trigonometric basis in the 2D continuum, and the spline basis and wavelet basis in the curve space. More details of the deformation basis can be found in [23].

We have used an *i.i.d.* zero-mean Gaussian distribution on the deformation parameters based on the assumption that the prototype represents the most likely object shape; the larger the deformation, the less likely the deformed template will be observed. The *i.i.d.* distribution is chosen for its simplicity although in reality the model parameters are rarely independent. To maximize the Bayesian *posterior* probability which combines both the prior knowledge of the deformation space and the image likelihood is equivalent to minimizing the following objective function [11]:

$$\mathcal{L}(\mathcal{T}_{s,\Theta,\underline{\xi},\underline{d}}, Y) = \mathcal{E}(\mathcal{T}_{s,\Theta,\underline{\xi},\underline{d}}, Y) + \gamma \sum_{\xi_i \in \underline{\xi}} |\xi_i|^2, \quad (2)$$

with respect to the deformation parameters $\underline{\xi}$ and the set of pose parameters (scale s , rotation Θ , and translation \underline{d}). The imaging term $\mathcal{E}(\mathcal{T}_{s,\Theta,\underline{\xi},\underline{d}}, Y)$ measures the agreement between the deformed template and the image [11]. It relates to the summation of potential energy of the pixels on the deformed template, where the potential field is computed based on the image features. The second term on the right hand side in (2) penalizes the deviation

from the prototype. The factor γ controls the “rigidness” of the template; it should be selected to reflect the variations in the object shape.

3 OBJECT TRACKING

In this section, we describe how to apply the prototype-based template model to track objects in an image sequence. We investigate all the possible cues that can be used to improve the tracking results. In particular, we use image gradient, interframe motion, and region correspondences to track the objects.

3.1 Tracking Criteria

Many object tracking applications share the following properties:

- 1) the interframe motion of the object is small so that the object in the next frame is in the neighborhood of the object in the current frame,
- 2) the same point on the object has a consistent color/gray scale in all the frames of the sequence,
- 3) the boundary of the moving object has a relatively large motion field, and
- 4) the boundary of the object has a large image gradient value.

Based on the above observations, we track the boundary of an object using the following criteria: 1) shape similarity: object shapes are similar in two successive frames, 2) region similarity: the properties (color, texture) of a region in the object remain constant throughout the sequence, 3) motion cue: the object boundary should be attracted to pixels with a large amount of motion, and 4) gradient cue: the object boundary should be attracted to pixels with a large image gradient. These criteria are explained in more detail in the following sections.

3.1.1 Matching Regions

Suppose the deformable template delineates the object boundary accurately in the first frame. The segmented region (object) is used as a reference object \mathcal{O}_{ref} for color and texture matching for the rest of the sequence, i.e., a point on the object in each frame exhibits similar region statistics as the corresponding point on the object in the first frame.

Suppose we have successfully tracked the object up to the i th frame. Since the interframe motion of the object is assumed to be small, the object boundary in the $(i+1)$ th frame is enclosed in a band (shaded region in frame $(i+1)$ in Fig. 1) centered at the object boundary in the i th frame. In other words, each boundary point in successive frames is enclosed in a disc centered at its current position. The radius of the disc depends on the interframe object motion. The larger the interframe motion, the larger the disc. When we track the object in the $(i+1)$ th frame, we can first predict

a radius for each boundary point based on the tracking result in the i th frame. Each point in this disc is compared to the object region around the corresponding boundary point in the reference object in terms of color or gray scale to compute a color or gray scale distance to the likely corresponding points on the reference object. Assume that the boundary of the object in the first frame consists of a linked list of n_0 points $p_0^0, p_1^0, p_2^0, \dots, p_{n_0-1}^0$, where the superscript indicates the frame number, and the subscript indicates which point it is on the object boundary. The neighborhood $\mathcal{N}^0(k)$ of the k th boundary point in the 0th frame is the intersection of a disc centered at p_k^0 and the object region. The neighborhood $N(l)$ of an image pixel l is defined to be the disc centered at l (see Fig. 1). For each point l in the band in the $(i+1)$ th frame, we compute a matching score which measures the color or gray scale similarity to possible corresponding points on the reference object. The RGB color space is used.

$$\text{distance}(l) = \min_{k, p_k^0 \in N(l)} \text{Dist}(l, \mathcal{N}^0(k)), \quad (3)$$

where $\text{Dist}(l, \mathcal{N}^0(k))$, the distance of pixel l to region $\mathcal{N}^0(k)$, is defined using the order statistic as follows: Let the Euclidean distances between the value of pixel l to the value of pixels in $\mathcal{N}^0(k)$ be $d_{lk_0}, d_{lk_1}, d_{lk_2}, \dots, d_{lk_{N_i-1}}$, such that $d_{lk_0} < d_{lk_1} < d_{lk_2} < \dots < d_{lk_{N_i-1}}$. $\text{Dist}(l, \mathcal{N}^0(k))$ is defined as the average of the distances between the 10th percentile and the 40th percentile, that is,

$$\text{Dist}(l, \mathcal{N}^0(k)) = \frac{\sum_{i=k_m}^{k_n} d_{lk_i}}{k_n - k_m + 1}, \quad (4)$$

where k_m and k_n are the 10th percentile and 40th percentile points. This statistic is used because it is robust to noise.

Given the converged deformable template in the i th frame, we can compute a distance map in the $(i+1)$ th frame, where for each point in the band around the template position in the i th frame, a distance is assigned to reflect its degree of resemblance to the potential object boundary. If this pixel happens to lie on the object boundary, the distance would be very small. If it is a background pixel which is quite different from the object pixels in color, the distance would be large. As a result, when we search for the possible object boundary in the new frame, it is not likely to be located in regions with large distance values.

3.1.2 Motion Cues

The motion field is obtained by computing the absolute values of the interframe differences and then smoothed using a 2D Gaussian mask.

3.1.3 Image Gradient

Object boundary (either static or moving) is often characterized by discontinuities in gray scale, which is indicated by a large image gradient. This criterion is most commonly used in image segmentation. We compute the image gradient for each image frame as the sum of squares of the greyscale differences along the x - and y - axes and then smooth it for each frame using a 2D Gaussian mask.

3.2 Objective Function

To track an object in an image sequence, we deform the template so that: 1) small deformations are preferred, 2) the template is attracted to image pixels with large gradient, 3) the template is attracted to image pixels with large motion (interframe distances),

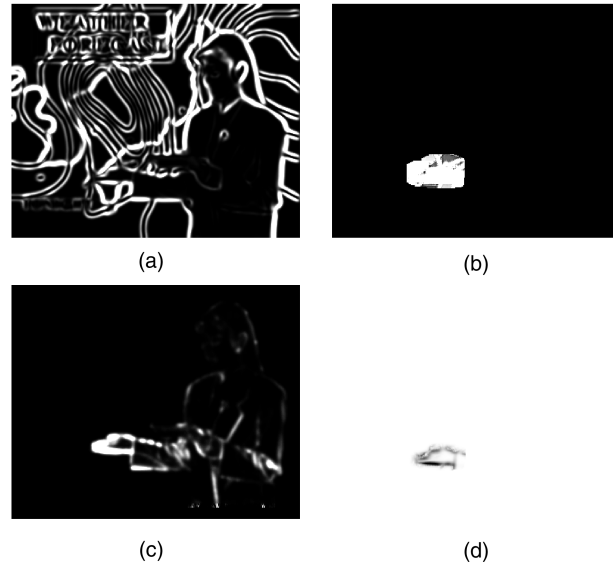


Fig. 2. Integrating image gradient, color consistency, and interframe motion. (a) Image gradient for an input frame, (b) color distance map (negated) for this frame, (c) interframe motion for this frame, and (d) integrated image potential field using (5).

and 4) the template deforms itself to minimize the average gray scale distance of the enclosed pixels.

The first goal is achieved by penalizing large deformation parameters. The remaining three goals are achieved using the following definition of image potential field.

3.2.1 Image Potential Field

We compute an image potential field to find a match between the deformable template and the object boundary in the image frame. The potential field for the tracking problem incorporates image gradient, gray scale cues, and motion cues. Let the image gradient plane for the i th frame be \mathcal{G}_i , the absolute difference between the $(i-1)$ th frame and the i th frame be \mathcal{M}_i , and the color distance map for the i th frame be \mathcal{C}_i , then the potential \mathcal{P}_i is computed as:

$$\mathcal{P}_i = -(\mathcal{G}_i \odot \mathcal{M}_i) \odot (\mathcal{C}_{max} - \mathcal{C}_i), \quad (5)$$

where \odot denotes pixelwise multiplication, and \mathcal{C}_{max} is the maximum color distance in \mathcal{C}_i .

The potential $\mathcal{P}_i(T)$ of a template T , when placed in the tracking potential field, is the average of potentials of all the template pixels. When the potential is minimized, the template is: 1) attracted to image edges, 2) attracted to moving boundaries, and 3) increasing its color (gray scale) similarity to the reference object.

The deviation of the template from the prototype is measured by the sum of squares of the deformation parameters. The objective function, which incorporates both the amount of deformation and the goodness of matching, is defined as:

$$\mathcal{L}(\xi, x, y, s) = \omega \left(\sum_{\xi \in \xi} \xi_i^2 \right) + \mathcal{P}(T). \quad (6)$$

We use the gradient descent method to minimize the objective function (6) *w.r.t.* the deformation parameters ξ , the translation parameters (x, y) and the scale parameter s .

Fig. 2 illustrates an example of the fusion of image gradient, color consistency, and motion. After the integration of the multiple image cues, an image potential field is obtained which highlights the desired salient image features.



Fig. 3. Tracking a human face in an image sequence (each frame size is 120×160). (a) Template initialization in the first frame and (b) tracking results for the sequence.

3.3 Tracking Algorithm

The tracking proceeds as follows:

1. **Initialization:** manually initialize a template at the proximity of the object in the *first* frame. Apply the deformable template matching process until it converges using gradient information only. Store the converged
2. **For all the frames from 2 to N :** Update the prototype to the converged deformable template in the previous frame, compute the gray scale or color distance map using the current frame and the reference object, compute the potential image by combining image gradient, interframe motion, and gray scale distance map depending on the

template and the gray scale or color information on the segmented object boundary and inside the region.



Fig. 4. Tracking a human hand in a weather forecast TV program (each frame size is 288×352).

problem at hand, initialize the template in the current frame using the pose (translation, scale) of the converged configuration in the previous frame, apply the deformable template matching process, and store the converged template for the current frame.

Note that although the deformation allowed between two consecutive frames is moderate, the change in shape through the sequence could be large due to the accumulation of the between-frame deformations.

4 EXPERIMENTAL RESULTS

We have applied the proposed tracking algorithm to track objects in a number of image sequences.

Human face tracking is of significant importance in a number of applications, including video conferencing where the face region is

located and coded at a higher bit rate than the background for transmission and storage. Fig. 3 shows the tracking results of a human face. The sequence consists of 35 frames (120×160 pixels) from a 12-second video segment, which was sampled at three frames per second. Image gradient and interframe motion are used to compute the image potential. Note that the template is capable of handling partial occlusions in the frames in the bottom row. Despite the shift and rotation of the face, the deformable template can reasonably delineate the contour of the head through the whole sequence. It takes 5.05 seconds (on a Sparc 20) to compute the potential image for the 35 frames, and another 2.53 seconds to track the face for the 12-second video clip.

Fig. 4 shows the tracking of a human hand in a weather forecast video clip using image gradient, color region constancy, and interframe motion. The map in the background contains curves

with very strong image gradient. The edge magnitude of the background curves is stronger than that at the boundary of the hand. However, with the help of color and motion information, we can reasonably track the moving hand through the 12 frames (each frame size is 288×352).

5 SUMMARY

We have presented an application of the prototype-based deformable template for tracking of objects in image sequences from different sources. The prototype-based deformable model has an advantage over the widely used "snake model" in tracking applications in that it inherently contains global structural information about the object shape, which makes it less sensitive to weak or missing image features.

We have combined image gradient, region color or gray scale and motion cues to facilitate the tracking process. In particular, we have introduced a region-based matching criterion which takes advantage of the color (gray scale) constancy of corresponding object pixels throughout the sequence. We have applied the algorithm to a number of image sequences from different sources. The experimental results are promising. The inherent structure in the deformable template, together with the region, motion, and image gradient cues, make the algorithm relatively insensitive to the adverse effects of weak image features and moderate degree of partial occlusion.

The proposed framework is quite general and can be applied to a number of tracking tasks. Future work will incorporate temporal prediction such as Kalman filtering to improve the tracking results. We will also investigate how to weigh the different image cues based on the uncertainty in estimating their values.

ACKNOWLEDGMENTS

This research was partially supported by a grant from Siemens Corporate Research, Princeton, New Jersey.

REFERENCES

- [1] N. Ayache, I. Cohen, and I. Herlin, "Medical Image Tracking," *Active Vision*, A. Blake and A. Yuille, eds., chapter 17, pp. 285-301, Cambridge, Mass.: MIT Press, 1992.
- [2] B. Bascle, P. Bouthemy, R. Deriche, and F. Meyer, "Tracking Complex Primitives in an Image Sequence," *Proc. 12th Int'l Conf. Pattern Recognition*, vol. 1, pp. 426-431, Oct. 1994.
- [3] B. Bascle and R. Deriche, "Region Tracking Through Image Sequences," *Proc. Fifth IEEE Int'l Conf. Computer Vision (ICCV)*, pp. 302-307, June 1995.
- [4] A. Blake, R. Curwen, and A.A. Zisserman, "A Framework for Spatiotemporal Control in the Tracking of Visual Contours," *Int'l J. Computer Vision*, vol. 11, no. 2, pp. 127-145, Oct. 1993.
- [5] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, "Active Shape Models—Their Training and Application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38-59, Jan. 1995.
- [6] D. DeCarlo and D. Metaxas, "The Integration of Optical Flow and Deformable Models: Applications to Human Face Shape and Motion Estimation," *Proc. IEEE Computer Vision and Pattern Recognition (CVPR '96)*, pp. 231-238, 1996.
- [7] P. Fieguth and D. Terzopoulos, "Color-Based Tracking of Heads and Other Mobile Objects at Video Frame Rates," *Proc. IEEE Conf. Computer Vision Pattern Recognition (CVPR '97)*, pp. 21-27, 1997.
- [8] P. Fua and Y.G. Leclerc, "Object-Centered Surface Reconstruction: Combining Multi-Image Stereo and Shading," *Int'l J. Computer Vision*, vol. 16, pp. 35-56, Sept. 1995.
- [9] B. Heisele, U. Krebel, and W. Ritter, "Tracking Nonrigid, Moving Objects Based on Color Cluster Flow," *Proc. IEEE Conf. Computer Vision Pattern Recognition (CVPR '97)*, pp. 257-260, 1997.
- [10] M. Isard and A. Blake, "Contour Tracking by Stochastic Propagation of Conditional Density," *Proc. European Conf. Computer Vision*, vol. 1, pp. 343-356, 1996.
- [11] A.K. Jain, Y. Zhong, and S. Lakshmanan, "Object Matching Using Deformable Templates," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 3, pp. 267-278, Mar. 1996.
- [12] I.A. Kakadiaris and D. Metaxas, "Model-Based Estimation of 3D Human Motion with Occlusion Based on Active Multiviewpoint Selection," *IEEE Computer Vision and Pattern Recognition*, pp. 81-87, 1996.
- [13] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active Contour Models," *Int'l J. Computer Vision*, vol. 1, no. 4, pp. 321-331, 1988.
- [14] C. Kervrann and F. Heitz, "Robust Tracking of Stochastic Deformable Models in Long Image Sequences," *Proc. Int'l Conf. Image Processing (ICIP)*, vol. 3, pp. 88-92, 1994.
- [15] A. Lanitis, C.J. Taylor, and T.F. Cootes, "A Unified Approach to Coding and Interpreting Faces," *Proc. Fifth IEEE Int'l Conf. Computer Vision (ICCV)*, pp. 368-373, 1995.
- [16] F. Leymarie and M. Levine, "Tracking Deformable Objects in the Plane Using an Active Contour Model," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 617-634, June 1993.
- [17] R. Malladi, J. Sethian, and B. Vemuri, "Shape Modeling with Front Propagation: A Level Set Approach," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 2, pp. 158-175, Feb. 1995.
- [18] N. Paragios and R. Deriche, "Geodesic Active Contours and Level Sets for the Detection and Tracking of Moving Objects," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 3, pp. 266-280, Mar. 2000.
- [19] D. Terzopoulos and R. Szeliski, "Tracking with Kalman Snakes," *Active Vision*, A. Blake and A. Yuille, eds., chapter 1, pp. 1-19, Cambridge, Mass.: MIT Press, 1992.
- [20] D. Terzopoulos, A. Witkin, and M. Kass, "Constraints on Deformable Models: Recovering 3D Shape and Nonrigid Motion," *Artificial Intelligence*, vol. 36, pp. 91-123, 1988.
- [21] D. Terzopoulos and K. Waters, "Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 569-579, June 1993.
- [22] A.L. Yuille and P.W. Hallinan, "Deformable Templates," A. Blake and A. Yuille, eds., *Active Vision*, MIT Press, 1992.
- [23] Y. Zhong, A.K. Jain, and M.P. Dubuisson-Jolly, "Object Tracking Using Deformable Templates," *Proc. Sixth IEEE Int'l Conf. Computer Vision (ICCV)*, pp. 440-445, 1998.
- [24] S.C. Zhu, T.S. Lee, and A.L. Yuille, "Region Competition: Unifying Snakes, Region Growing, Energy/Bayes/MDL for Multiband Image Segmentation," *Proc. Fifth IEEE Int'l Conf. Computer Vision (ICCV)*, pp. 416-423, 1995.