43rd IEEE Conference on Decision and Control
December 14-17, 2004
Atlantis, Paradise Island, Bahamas

Tracking Deformable Moving Objects Under Severe Occlusions

Jeremy D. Jackson Anthony J. Yezzi Stefano Soatto

Abstract— **We propose a nonlinear model for tracking a slowly deforming and moving contour despite significant occlusions. The contour is represented implicitly, and its motion is described by the action of a finite-dimensional group; we estimate both the implicit representation of the contour (its shape) and its motion. Our contribution consists in defining a generative model that is not subject to arbitrary re-parameterization, choice of (non-unique) key points or control points, and allows enforcing a dynamical model of motion when it is available. Otherwise, our approach allows enforcing simple phenomenological models, for instance low acceleration or low jerk.**

I. Introduction

We are interested in tracking moving objects in a sequence of images. For us, an "object" is a region of the image that has a distinct photometric signature, something that distinguishes it from the rest of the image, or "background." For instance, it could have quasi-homogeneous intensity, or some other statistic that is uniform or almost uniform within the object, but distinct from the rest of the image ... for the most part. In particular, we are interested in being able to track the object despite it being invisible or partly invisible at certain instants of time. In addition, we want to be able to track the object despite changes in the geometry, and possibly the topology, of the region that determines it.

This latter issue of deformations has received significant attention in the literature, which we review briefly in the next subsection. In particular, [29] characterizes the tracking of a moving object, where motion can be defined by a finite-dimensional group (for instance affine), through the introduction of a generative model of the so-called "average shape," from which each measurement is obtained with minimum deformations, measured with respect to a chosen energy functional. While that work hinted at the problem of extending the framework to the case when the underlying shape average is changing over time, it did not offer a technical solution for tracking with an explicit dynamical model.

Now, in order to track regions through occlusions, a motion model is necessary to predict, or extrapolate, the state of the object in lack of measurements. This issue of learning motion dynamics has also gathered considerable attention in the past, and indeed some of the most popular particle filtering techniques were developed in the context of contour tracking [6]. In this paper, however, we consider objects as regions bounded by closed planar contours, represented implicitly. These are infinite-dimensional objects, and there is no existing filtering technique suitable for such infinite-dimensional state-spaces. Therefore, this manuscript represents a small step in a significantly novel and challenging direction, as we describe in the next subsection. Before we do so, however, we point out that – although a rigorous solution to this problem is elusive – approximate filtering can be performed in a way that results in simple, robust and efficient algorithms that we validate experimentally in Section IV on sequences of images with severe occlusions.

A. State of the art and our contribution

Contour tracking has been a very active area of research in vision for many years. The book of Blake and Isard contains a snapshot of the state of the art as of 1998 [6]. What makes the problem different from a standard tracking problem, as studied in signal processing and control theory, is the fact that the representation of the state of the model and of the measurement map is non-trivial, whereas in traditional tracking the "target" is usually a point or a collection of points in a vector space. In particular, a common approach in contour tracking is to represent the contour using a finite-dimensional representation. This includes various types of splines or "snakes" (see [6], [26] and references therein), various discretizations of the contour, for instance using polygonal approximations [27], [28]. In all these finite-dimensional approximations, a dynamical model is introduced by modeling the *parameters* (e.g. the nodal points, or the control points, of the representation) as the state of a dynamical model, typically assumed linear (e.g. autoregressive moving-average model). The difficulty with this approach is that each contour is represented *not* by a set of parameters, but by an entire equivalence class of parameters obtained by moving the control/nodal points: There are infinitely many choices of control points that result in substantially the same measured contour. Therefore, many have resorted to additional constraints, for instance equi-spacing of polygonal vertices, fixed

J. D. Jackson and A. J. Yezzi are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 gtg120d@mail.gatech.edu, ayezzi@ece.gatech.edu.

S. Soatto is with the Computer Science Department, University of California at Los Angeles, Los Angeles, CA 90095 soatto@ucla.edu.

Supported in part by AFOSR E-16-V91-G2/F49620-03-1-0095.

number of equi-distant control points etc. Additional techniques rely on describing regions using "blobs" or other objects with pre-specified shape, or by collections of spatial configurations of blobs. This is common for the case of cars and people (see for instance [7], [4] and references therein); such techniques have proven successful even in the case of severe occlusions [11].

A substantially different approach is taken when the contour is represented in the continuum. For instance, "deformable templates," pioneered by Grenander [10], do not rely on "features" or "landmarks;" rather, images are directly deformed by a (possibly infinite-dimensional) group action and compared for the best match in an "image-based" approach [31], [2], [30], [20], [16], [25], [17], [15], [9], [18]. A common model is to represent the contour implicitly, as the zero level set of a function (e.g. a signed distance function) that evolves in time. For instance, geodesic active contours [8], [14] have been successfully used for tracking, for instance, cardiac motion in ultrasound imaging [24]. In most of the current approaches, however, "time" only indicates the index of the iterative procedure used to estimate the contour, and most motion models are essentially assuming that the position of the object at time $t+1$ is close to that at time t, and therefore the best estimate of the contour at time t can be used to initialize the same procedure at the next instant [21]. In this paper, we want to be able to enforce higher-order motion models, for instance due to inertia and other constraints on acceleration. The idea is to set up a framework where a detailed motion model can be used, if available, and other statistical or phenomenological motion models can be used otherwise. For instance, we may want to enforce regularity by imposing that velocity is small, or that acceleration is small, or that jerk (the derivative of acceleration) is small and so on.

Ideally, we would like to derive an optimal framework to do so. This would entail estimating the conditional density of the state (motion and shape of the deforming contour) given the measurements up to time t (noisy/deformed measurements of the contour, possibly with significant missing pieces). This is easy to do for linear dynamical models driven by additive, white, zero-mean Gaussian noise, but is out of the question for a state that is infinite-dimensional (the shape of the deforming contour), has non-trivial geometry (the group structure), highly non-linear measurement equations (due to occlusions), and the uncertainty is functional, rather than additive (the diffeomorphic model of the contour deformation). Therefore, we can only resort to approximate filtering techniques, with no available analytical statements about their performance. While filtering for non-linear finite-dimensional models has received a lot of attention since the age of Wiener in the mid forties, and has culminated in several viable, although not-proven-optimal techniques, such as the Extended Kalman Filter [12], the multimodal sum-of-Gaussian filter [1], particle filtering [6], various forms of multi-modal, multi-target tracking based on interacting multiple models (see [3] and references therein) and various numerical approximations of the Mortensen-Zakai equation, there is very little work on filtering for infinite-dimensional state spaces. Blake and Brockett first confronted this problem in [5], where they address the problem of estimating a moving curve (represented as the graph of a function) despite missing segments of the curve. In our case the problem is more difficult because we cannot rely on the graph structure, and furthermore our solution is entirely different from that suggested in [5].

II. Formalization of the problem

At any instant of time $t \in \mathbb{R}$, let $\mu(t) : \mathbb{S}^1 \to \mathbb{R}^2$ be a closed planar contour, $g(t) \in G$ be a finite-dimensional group action (e.g. the Euclidean group $G = SE(2)$ or the affine group $G = A(2)$), and $h_t : \mathbb{R}^2 \to \mathbb{R}^2$ a diffeomorphism that can change over time (hence the subscript t). We measure a closed planar contour $y(t) : \mathbb{S}^1 \to \mathbb{R}^2$ that is a local deformation h_t of the static contour μ moving under the action of g. Therefore, formally we can write a generative model for the data $y(t)$ as follows:

$$\begin{cases} \dot{\mu}(t) = 0 \\ \dot{g}(t) = \widehat{v} g(t) \\ \dot{v}(t) = \alpha(t) \\ y(t) = h_t(g(t)\mu(t)). \end{cases} \tag{1}$$

In this model, the first equation embodies the assumption that the average shape is constant. If this is not the case, but still it varies slowly relative to the intrinsic dynamics of $y(t)$, we could write formally that $\dot{\mu}(t) = w(t)$. We will comment later on what this notation actually means. If there are no assumptions made on how the average shape evolves, the tracking problem cannot be meaningfully addressed [29]. The second equation is just a deterministic integrator that says that pose is the integral of velocity, and the third equation says that velocity is the integral of acceleration, which we do not know. We could assume, again, that α is an unknown constant, or that it varies slowly relative to the dynamics of $y(t)$. Finally, the last equation says that the measurements are a perturbation of the average shape in the moving frame. Our goal is to infer μ, g and v from measurements of y. In particular, we are interested in the estimate that results in the "smallest" possible perturbation $h_t(\cdot)$. This model is just formal notation,

and in order to design, implement and analyze inference algorithms we must specify (i) a representation for μ, (ii) a local coordinate system for g, and (iii) a discrepancy measure between the data $y(t)$ and the model of the data, $h_t(g(t)\mu(t))$. The second issue is straightforward since canonical coordinate charts for matrix Lie groups are easy to derive and compute using the exponential map [19]. The first and the third issue are more complex and highly interconnected.

In fact, consider a parametric representation of $\mu(t)$, for instance $s \mapsto \tilde{\mu}(t,s)$. The measured contour $y(t)$ can also be parameterized via $l \mapsto \tilde{y}(t,l)$. Unfortunately, the *correspondence* of s and l is not known, and therefore the measurement equation relies on an estimate of the reparametrization $l \mapsto s = \rho(l)$, or on a canonical representative of the equivalence class. This significantly complicates the model, since we now have $y(t) = \tilde{y}(t,l) = h(g(t)\tilde{\mu}(t,\rho(l)))$, and we have no constraints on ρ other than it being a smooth bijection. Therefore, we choose to represent μ *implicitly* as a set $\mu(t) = \{x \in \mathbb{R}^2 \mid \chi_\mu(x,t) \leq 0\}$, where χ denotes the characteristic function of the set μ. While this causes no problem for the basic model where $\dot{\mu} = 0$, in the case of "slowly varying" average shape, we have to define what $w(t) \doteq \dot{\mu}(t)$ means. Furthermore, in lack of a correspondence $s \leftrightarrow l$, we need to specify how we compute a discrepancy between $y(t)$ and $h_t(g(t)\mu(t))$. For now, we will simply indicate with $w(t)$ the quantity defined by the equation $\mu(t+dt) = \mu(t) \oplus w(t)$, where $\oplus$ denotes a composition operation in the set $\mu(t)$, for instance $w(t)$ can be the set-symmetric difference between $\mu(t+dt)$ and $\mu(t)$.

III. Filtering deforming shapes

With this formalism, we can now postulate the structure of the state estimator. We start at $t = 0$ with an initial point estimate, $\{\hat{\mu}(0), \hat{g}(0), \hat{v}(0)\}$. Since the global reference is arbitrary, we typically choose $\hat{g}(0) = Id$, the group identity. Now, at time t, in lack of any new measurement, the best estimate of the state at $t+\delta$ is obtained by integrating the state equation between t and $t+\delta$. We will choose δ to be the unit of time; integrating the basic model for constant μ and constant velocity $\alpha = 0$ is straightforward, since we have that $\hat{\mu}(t+1) = \hat{\mu}(t), \hat{v}(t+1) = \hat{v}(t)$ and $\hat{g}(t+1) = \exp(\widehat{v}(t))\hat{g}(t)$, where "$\widehat{\ }$" is the operator onto the Lie algebra, and we have omitted the superscript "$\widehat{\ }$" from $v(t)$ for ease of notation. Therefore, the prediction step is trivial.

Now, assuming that a new measurement $y(t+1)$ becomes available, we are interested in updating the prediction in a way that guarantees that the estimate of the state will converge, asymptotically, to the true state. While in the case of a linear finite-dimensional model one can derive the optimal estimator directly, here in general there is no finite-dimensional optimal estimator. Therefore, we will postulate a structure of the estimator, in the form of a generic local observer, and then choose the parameters that guarantee error stability.

Since the deterministic integrator $\dot{g} = \widehat{v}g$ is not imposing any model constraint (other than adherence to the group action G), that equation carries no uncertainty: if v was known exactly and g was known exactly, then $\dot{g}$ would be given *exactly* by $\widehat{v}g$. Therefore, that equation carries no model error and the filter can be saturated along the corresponding components. In practice, we write that equation in local coordinates Ω, defined by $g \doteq e^{\widehat{\Omega}}$, and approximate the equation to first order as $\dot{\Omega} = (I + \widehat{\Omega})v$. Consequently, the measurement equation becomes, neglecting the time indices, $y = h(e^{\widehat{\Omega}}\mu)$.

The goal of the update step is to reduce the "uncertainty," i.e. the discrepancy of the model from the measurements. Since in our case the uncertainty is the diffeomorphism h_t, not your usual additive noise, at each step we have to solve a local optimization in order to compute the best update for the state. In particular, we will consider a local update based on an incremental step in the direction of the gradient of a cost functional that measures the "energy" of the diffeomorphism h_t, subject to the model (1). Specifically, at time t we consider a causal window of length $k \geq 2$, and look at the energy

$$\phi(v_1, \ldots, v_k, \Omega, \mu) \doteq \sum_{\tau=t-k+1}^{t+1} \int E(h_\tau(x))dx \qquad (2)$$

subject to $y(\tau) = h_\tau\left(e^{\widehat{v}_1} \ldots e^{\widehat{v}_k} e^{\widehat{\Omega}(\tau-k)}\mu(\tau)\right)$.

Now, to quantify the discrepancy between the model and the measurements, indicated by $E(h_t)$ above, we utilize a discrepancy function *inside* the region $g\mu \subset \mathbb{R}^2$, f_{in}, and a discrepancy function *outside* the region, f_{out}. These can be as simple as the indicator functions $f_{in}(x,y) = \chi_y(x) - 1$, and $f_{out}(x,y) = \chi_y(x)$ for the case of binary images representing evolving shapes, or more complex signed-distance scores that can be generalized to grey-scale images as it has been shown in [29]. In either case, we write

$$E(h_\tau) = \int_{g(\tau)\mu(\tau)} f_{in}(x, y(\tau))dx + \int_{g(\tau)\mu^c(\tau)} f_{out}(x, y(\tau))dx. \qquad (3)$$

We report the computation of the gradients $\nabla_\mu \phi$, $\nabla_\Omega \phi$ in the next section. Once these gradients have been computed numerically, the general form of the update

becomes

$$\begin{cases} \hat{\mu}(t+1) = \hat{\mu}(t) \oplus L_\mu \nabla_\mu \phi\left(\hat{v}(t), \ldots, \hat{v}(t-k), \hat{\mu}(t)\right) \\ \hat{g}(t+1) = e^{\widehat{v}(t)} \hat{g}(t) \\ \hat{v}(t+1) = \hat{v}(t) + L_v \nabla_v \phi\left(\hat{v}(t), \ldots, \hat{v}(t-k), \hat{\mu}(t)\right) \end{cases} \tag{4}$$

where L_μ, L_v are tuning parameters whose effects are discussed in the next section. In the initialization phase, rather than running one step of the gradient above, we run several until convergence to steady state of ϕ. Furthermore, depending on the convergence rate of ϕ relative to the dynamics of μ, it is useful to run several steps of the gradient, or even to steady state, in the update equation above.

IV. Experiments

In our implementation we have used a numerical computation of the gradient above to generate an update for the evolving shape represented implicitly within the level set framework [23]. In particular,

$$\nabla_\mu \phi = \sum_{\tau=1}^{k} |J_g(\tau)|(f_{in}(g(\tau)x, y(\tau)) - f_{out}(g(\tau)x, y(\tau)))N \tag{5}$$

where N is the outward unit normal and J_g is the Jacobian of g. The update equation for g, or equivalently Ω, is just the integrator described in the previous section. To update v, we update each component v_i independently using

$$\nabla_{v_i} \phi = \sum_{\tau=1}^{n} \int_{g^{-1}(\tau)y(\tau)} \langle \frac{\partial g(\tau)x}{\partial v_i}, f_{\{in,out\}}(g(\tau)x, y(\tau)) J_{g_* T}(\tau) \rangle ds \tag{6}$$

where g_* denotes the push-forward and T is the unit tangent vector,

Varying the gain L_μ one enforces more or less inertia by μ to change shape. In Figure 1, a vertical bar five pixels wide has been removed from the images to create an occlusion. The occlusion is close in grayscale value to that of the person being tracked. To see the effect of varying the gain on the estimation of the contour as it passes behind the occlusion, we use an image sequence with a model of motion that is fairly simple to track (constant velocity) and therefore use a low gain on the registration(motion) parameters. As the person passes under the occlusion, the contour will be attracted to the occlusion and without some state estimator it would grab onto the occlusion. So in Figure 1, we vary $L_\mu = 0.1, 0.65, 0.7$. $L_\mu = 0.65$ experimentally turns out to segment the person the best while avoiding getting caught up on the occlusion.

Figure 2 has a total occlusion. Here again the model of motion is certain (constant acceleration) and a low gain on the motion parameters is used. The gains chosen for the very uncertain contour are $L_\mu = 0.1, 0.5, 0.8$. The initial gain for the contour is set quite low so measurements are still a bit emphasized, but then the model is able to take over and push the tracking of the person through the total occlusion. $L_\mu = 0.1$ tracks the person but the contour is very rigid. For $L_\mu = 0.5$, the person is tracked but the contour gets thrown off by the similar looking books on the printer and poster on the wall. The last example $L_\mu = 0.8$ emphasizes the measurements too much and loses the person.

In Figure 3, there is only a slight occlusion as the car passes underneath the light pole. But the occlusion is very different from the car. The gain L needs to be chosen a bit higher than the previous examples because the car is going into a turn. Now there is uncertainty in the model of motion (constant acceleration) and the segmentation. A low gain ($L=0.1$ case) on all of the motion parameters and the contour would emphasize the model and would keep the contour tracking in the original direction the car was moving. $L = 0.1, 0.3, 0.7$ are looked at in this example. The contour is only slightly affected as it passes underneath the pole in the case of $L = 0.3, 0.7$.

V. Conclusions

We have presented a first step in designing a filter for a dynamical model of an evolving contour. The contour is represented implicitly as the (infinite-dimensional) locus of zeros of a given function, that evolves in time under the action of a group. We estimate both the underlying state and the group action, from noisy images that can have significant portions of missing data. Although it is hard to say anything analytically about the behaviour of such a filter, we have experimented with challenging real sequences and we have obtained very encouraging results.

References

[1] D. Alspach and H. Sorenson. Nonlinear bayesian estimation using gaussian sum approximations. *IEEE Trans. on Automatic Control*, 17:439–448, 1972.

[2] R. Azencott, F. Coldefy, and L. Younes. A distance for elastic matching in object recognition. *Proc. 13th Intl. Conf. on Patt. Recog*, 1:687–691, 1996.

[3] Y. Bar-Shalom and T. E. Fortmann. *Tracking and Data Association*. Academic Press, 1987.

[4] M. J. Black. Eigentracking: robust matching and tracking of articulated objects. 1996.

[5] A. Blake and R. Brockett. On snakes and estimation theory. *Submitted to the invited session on "Dynamic Vision" at the 33 Conf. on Decision and Control*, 1994.

[6] A. Blake and M. Isard. *Active contours*. Springer Verlag, 1998.

[7] C. Bregler. Learning and recognizing human dynamics in video sequences. In *Proc. of the Conference on Computer Vision and Pattern Recognition*, pages 568–574, 1997.

Fig. 1. Tracking a Person through an Attracive Occlusion, Top Row: $L_\mu = 0.1$, Middle Row: $L_\mu = 0.65$, Bottom Row: $L_\mu = 0.7$

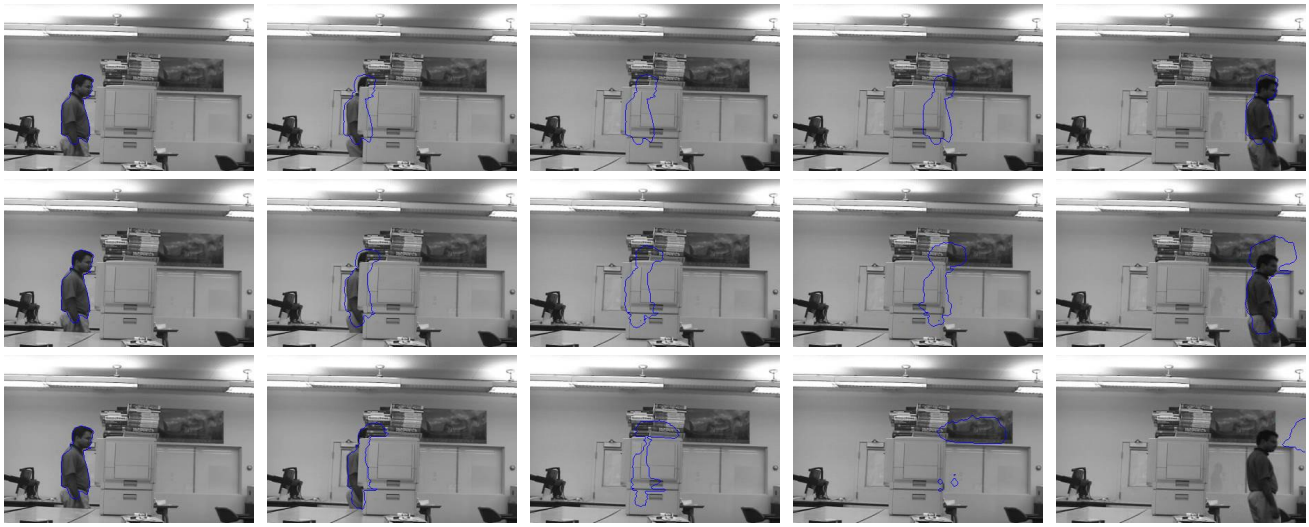

Fig. 2. Tracking a Person through a Severe Occlusion, Top Row: $L_\mu = 0.1$, Middle Row: $L_\mu = 0.5$, Bottom Row: $L_\mu = 0.8$

[8] V. Caselles, R. Kimmel, G. Sapiro, "Geodesic active contours," *Proceedings of the ICCV*, 1995.

[9] T. Chan and L. Vese. An active contours model without edges. In *Int. Conf. Scale-Space Theories in Computer Vision*, pages 141–151, 1999.

[10] U. Grenander. *General Pattern Theory*. Oxford University Press, 1993.

[11] C. Rasmussen and G. Hager. Probabilistic Data Association Methods for Tracking Complex Visual Objects *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:560–576, 2001.

[12] A.H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, 1970.

[13] H. Khalil, *Nonlinear Systems*, MacMillan Publ. House, N.J, 1992

[14] S. Kichenassamy, A. Kumar, P. Olver, A. Tannenbaum, and A. Yezzi, "Gradient flows and geometric active contours," *Proceedings of ICCV*, Cambridge, MA., June 1995.

[15] B. Kimia, A. Tannebaum, and S. Zucker. Shapes, shocks, and deformations i: the components of two-dimensional shape and the reaction-diffusion space. *Int'l J. Computer Vision*, 15:189–224, 1995.

[16] R. Kimmel and A. Bruckstein. Tracking level sets by level sets: a method for solving the shape from shading problem. *Computer Vision, Graphics and Image Understanding*, (62)1:47–58, 1995.

[17] R. Kimmel, N. Kiryati, and A. M. Bruckstein. Multivalued distance maps for motion planning on surfaces with moving obstacles. *IEEE Trans. Robot. & Autom.*, 14(3):427–435, 1998.

[18] M. Leventon, E. Grimson, and O. Faugeras. Statistical shape influence in geodesic active contours, 2000.

[19] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *An invitation to 3D vision, from images to models*. Springer Verlag, 2003.

[20] M. I. Miller and L. Younes. Group action, diffeomorphism and matching: a general framework. In *Proc. of SCTV*, 1999.

[21] M. Moelich and T. Chan. Tracking objects with the Chan-Vese

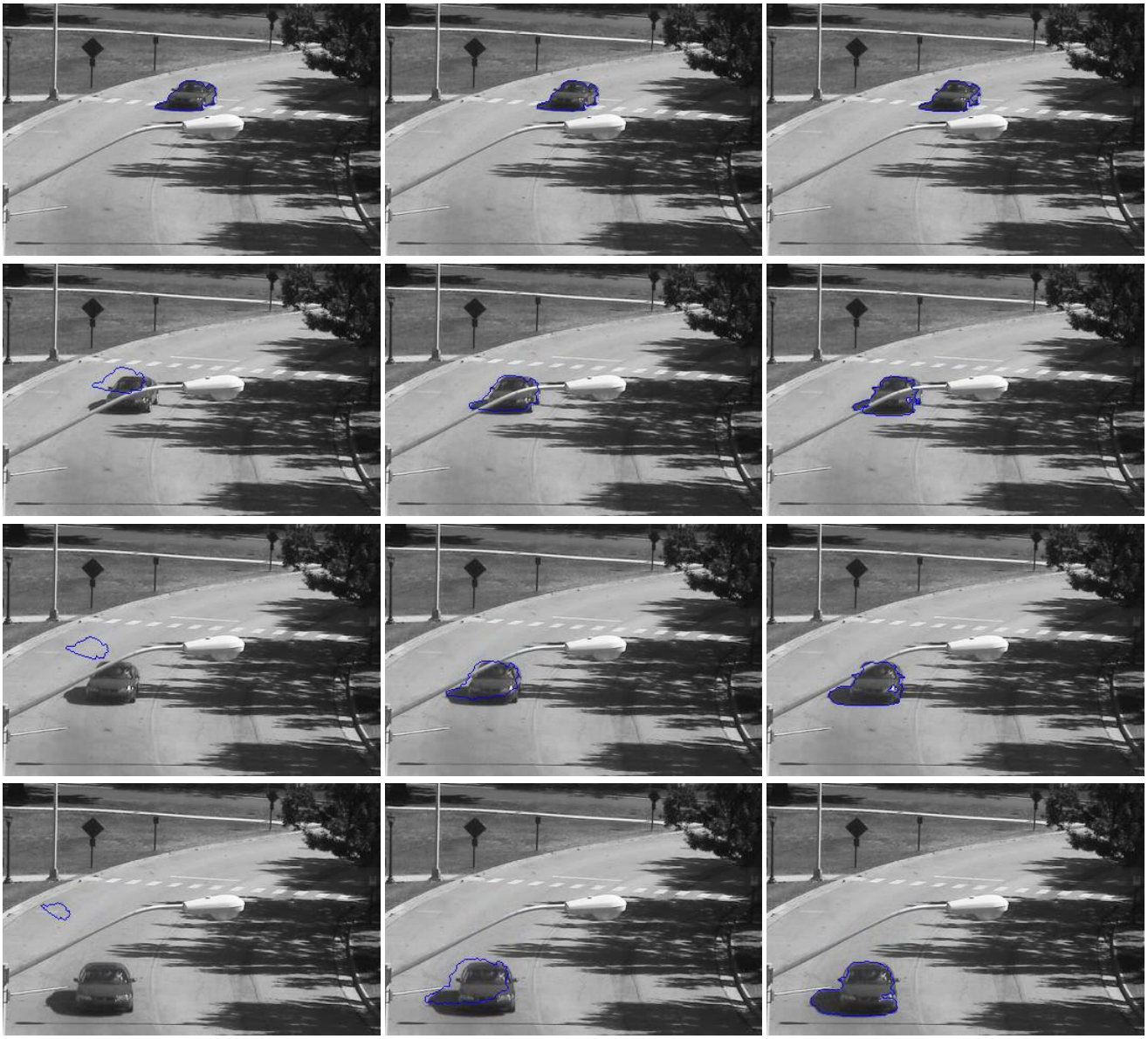

Fig. 3. Tracking a Car Under a Non-Attractive Occlusion, Left Column: $L = 0.1$, Middle Column: $L = 0.3$, Right Column: $L = 0.7$

model. UCLA CAM Report 3-14, March 2003.

[22] Mumford D., Shah J.: Optimal approximations by piecewise smooth functions and associated variational problems. In *Commun. Pure Appl. Math (1989)*, vol. 42, no.4, pages 577-684, 1989.

[23] S. Osher and J. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi equations. *J. of Comp. Physics*, 79:12–49, 1988.

[24] N. Paragios and R. Deriche. Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(3):266–280, 2000.

[25] C. Samson, L. Blanc-Feraud, G. Aubert, and J. Zerubia. A level set model for image classification. In *in International Conference on Scale-Space Theories in Computer Vision*, pages 306–317, 1999.

[26] M. Kass, A. Witkin, and D. Terzopoulosi. Snakes: active contour models. In *First International Conference on Computer Vision*, pp. 259–268, 1987.

[27] G. Unal, A. Yezzi, and H. Krim, "A Vertex-based Representation of Objects in an Image," *Proc. of Int. Conf. Image Processing*, vol. 1, September 2002, pp. 896–899.

[28] G. Unal, A. Yezzi, and H. Krim, "Active Polygons for Object Tracking," *First International Symposium 3D Data Processing, Visualization, Transmission*, Padova,Italy, June 2002.

[29] A. Yezzi and S. Soatto. Deformotion: deforming motion, shape average and the joint segmentation and registration of images. *Intl. J. of Comp. Vis.*, 53(2):153–167, 2003.

[30] L. Younes. Computable elastic distances between shapes. *SIAM J. of Appl. Math.*, 1998.

[31] A. Yuille. Deformable templates for face recognition. *J. of Cognitive Neurosci.*, 3(1):59–70, 1991.