In *Computer Vision and Pattern Recognition (CVPR 99)*, Ft. Collins, CO, pages 239-245, June, 1999.

A Multiple Hypothesis Approach to Figure Tracking

Tat-Jen Cham James M. Rehg
Cambridge Research Laboratory
Compaq Computer Corporation
Cambridge MA 02139
tjc@crl.dec.com rehg@crl.dec.com

Abstract

This paper describes a probabilistic multiple-hypothesis framework for tracking highly articulated objects. In this framework, the probability density of the tracker state is represented as a set of modes with piecewise Gaussians characterizing the neighborhood around these modes. The temporal evolution of the probability density is achieved through sampling from the prior distribution, followed by local optimization of the sample positions to obtain updated modes. This method of generating hypotheses from state-space search does not require the use of discrete features unlike classical multiple-hypothesis tracking. The parametric form of the model is suited for high-dimensional state-spaces which cannot be efficiently modeled using non-parametric approaches. Results are shown for tracking Fred Astaire in a movie dance sequence.

1 Introduction

Visual tracking of human motion is a key technology in a large number of areas. It has applications ranging from 3D mouse input [1] to content-based video editing [2]. This paper addresses the visual tracking problem for an articulated object such as the human figure, using a known kinematic model [3, 4, 5, 6]. The kinematics of an articulated object provide the most fundamental constraint on its motion. Kinematic models play two roles in tracking. First, they define the desired output—a state vector of joint angles that encodes the degrees of freedom of the model. Second, they specify the mapping between states and image features that makes registration possible.

A key attribute of any tracking scheme is the choice of probabilistic representation for the state estimates. The Kalman filter [7] is a classical choice which has been employed in earlier figure tracking work (see [8, 9, 10] for examples). Unfortunately the Kalman filter is restricted to representing unimodal probability distributions. The presence of background clutter, self-occlusions, and complex dynamics during figure tracking results in a state space density function (pdf) which is multi-modal.

Multiple hypothesis tracking (MHT) is a classical approach to representing multimodal distributions with Kalman filters [11]. It has been used with great effectiveness in radar tracking systems, for example. This method maintains a bank of Kalman filters, where each filter corresponds to a specific hypothesis about the target set. In the usual approach, each hypothesis correspond to a postulated association between the target and a measured feature. The multiple hypotheses arise when there are two or more features for which the correct association is not known. These methods however assume that a set of discrete features can be obtained at each time step, which presupposes that such a sensor exists. This is often not true when tracking complex objects – for example, there is no simple detector for the human figure which takes an input image and explicitly returns 'figure features' where each feature specifies a different skeletal configuration.

One alternative is to use Monte Carlo methods such as Isard and Blake's CONDENSATION algorithm [12]. While nonparametric models can represent arbitrary pdfs, their computational costs are prohibitive for the large state spaces required in figure tracking.

This paper describes a novel formulation of MHT for figure tracking. The key idea is to explicitly model and track the modes in the state pdf. We use a sampling-based state space search process to generate a set of hypotheses corresponding to the local maxima in the likelihood. By generating hypotheses through state space search we avoid the need for a complex figure detector necessary to apply classical MHT methods. By explicitly focusing our representation on the modes of the distribution we avoid the explosion in the number of samples that a Monte-Carlo-based scheme requires. A more detailed comparison between our proposed formulation and these methods is made in section 5.1. Our approach is based on the observation that complex targets such as the human figure usually have only small number of well-defined minima in their posterior density.

This work is the first application of multiple hypothesis techniques to figure tracking. An earlier version of this paper may be found in [13]. A more detailed analysis is also provided in [14].

1.1 The 2D Scaled Prismatic Model

Much of the previous work on figure tracking has employed 3D kinematic models and focused on detailed estimation of 3D motion. These approaches require multiple camera viewpoints for accurate estimation and rarely operate on-line. In contrast, perceptual user interface applications are more likely to benefit from reliable 2D figure tracking that can operate in real-time using a single camera input. For example, it's likely that many useful gestures can be recognized from a purely image-based description of figure motion, without recourse to 3D motion estimates.

This paper focuses on figure registration, which is the estimation of 2D image plane figure motion across a video sequence. Figures are described by a novel class of 2D kinematic models called *Scaled Prismatic Models* (SPM), introduced in [2]. These models enforce 2D constraints on figure motion that are consistent with an underlying 3D kinematic model. Unlike 3D kinematic models, SPM's do not require detailed prior knowledge of figure geometry and do not suffer from singularity problems when they are used with a single video source.

Each link in a scaled prismatic model describes the image plane appearance of an associated rigid link in an underlying 3D kinematic chain. Each SPM link can rotate and translate in the image plane, as illustrated in Figure 1. The link rotates at its joint center around an axis which is perpendicular to the image plane. This captures the effect on link orientation of an arbitrary number of revolute joints in the 3D model. The translational degree of freedom (DOF) models the distance between the joint centers of adjacent links. It captures the foreshortening that occurs when 3D links rotate into and out of the image plane. This DOF is called a scaled prismatic joint because in addition to translating the joint centers it also scales a template representation of the link appearance.

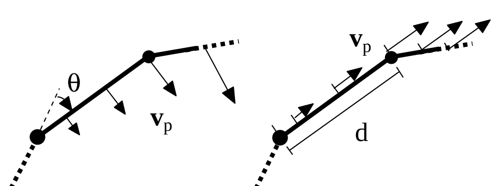


Figure 1: The effect of revolute (θ) and prismatic (d) DOF's on one link from a 2D SPM chain. The arrows show the instantaneous velocity of points along the link due to an instantaneous state change.

A complete discussion of SPM models, including a derivation of the SPM Jacobian and an analysis of its singularities, can be found in [2]. In this report we model the figure as a branched SPM chain. Each link in the arms, legs, and head is modeled as an SPM link. Each link has two degrees of freedom, leading to a total body model with 19 DOF's. The tracking problem consists of estimating a vector of SPM parameters for the figure in each frame of a video sequence, given some initial state.

2 Probability Density Representation

The choice of representation for the probability density of a tracker state is largely dominated by two concerns. The unimodality constraint imposed when using a Gaussian-based parametric representation such as the Kalman Filter is inaccurate when tracking in a cluttered environment, while a sample-based representation (such as used in the CONDENSATION algorithm) requires a prohibitive number of samples for encoding the probability distribution of a high-DOF SPM model. Instead we adopt a hybrid representation which supports a multimodal description but requires fewer samples for modeling.

Our selected representation is based on retaining only the modes (or peaks) of the probability density and modeling the local neighborhood surrounding each mode with a Gaussian. This addresses the multimodality issue directly, while the use of Gaussians eliminates the need for a large number of samples to non-parametrically shape the distribution around each mode.

3 Mode-based Multiple-Hypothesis Tracking

The basic idea in a probabilistic framework for tracking involves maintaining a time-evolving probability distribution of the tracker state. In order to generate a mode-based representation for the probability distribution of the tracker state, the algorithm has to recover these modes in each time-frame.

The algorithm proposed here may be modularized in a manner compatible with Bayes Rule:

$$p(x_t|Z_t) = k\; p(z_t|x_t)\; p(x_t|Z_{t-1}) \tag{1}$$

where x_t is the tracker state at time t, z_t is the observed data, Z_t is the aggregation of past image observations (ie. z_τ for $\tau = 0, \ldots, t$), and k is a normalization constant. Furthermore z_t is assumed to be conditionally independent of Z_{t-1} given x_t.

The stages of the algorithm at each time-frame are

1. Generating the new prior density $p(x_t|Z_{t-1})$ by passing the modes of $p(x_{t-1}|Z_{t-1})$ through the Kalman filter prediction step.

2. Likelihood computation, involving:

   (a) Creating initial hypothesis seeds by sampling the distribution of $p(x_t|Z_{t-1})$.

(b) Refining the hypotheses through differential state-space search to obtain the *modes* of the likelihood $p(z_t|x_t)$.

(c) Measure the local statistics associated with each likelihood mode using perturbation analysis.

3. Computing the posterior density $p(x_t|Z_t)$ via Baye's Rule (1), then updating and selecting the set of modes.

3.1 Multiple Modes as Piecewise Gaussians

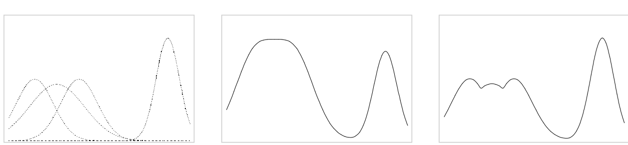
Given a set of N modes for which the ith mode has a state $\boldsymbol{m}_i$, an estimated covariance $\boldsymbol{S}_i$ and a probability p_i, an accurate construction of the probability density function requires a local maxima of value p_i located at each $\boldsymbol{m}_i$, with the local neighborhood surrounding $\boldsymbol{m}_i$ being approximately Gaussian with covariance $\boldsymbol{S}_i$.

In situations when the modes can occur in clusters (as is often the case), it is erroneous to use the individual modes directly as components in a *Gaussian sum* representation. Consider the simplified example for four hypotheses in 1D state-space as shown in fig. 2(a). If the hypotheses are directly considered the components in a Gaussian sum, the combined pdf has only two modes. This is shown in fig. 2(b). This results in a cluster of weaker modes being over-represented at the expense of strong but isolated modes. Instead we propose a Piecewise Gaussian (PWG) representation where the probability density $p(\boldsymbol{x})$ at a point $\boldsymbol{x}$ in the state-space is determined by the Gaussian component providing the largest contribution at $\boldsymbol{x}$, ie.

$$p(\boldsymbol{x}) = k \max_{i=1..N} \left\{ p_i \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_i)^T \boldsymbol{S}_i^{-1}(\boldsymbol{x}-\boldsymbol{m}_i) \right) \right\} \quad (2)$$

where k is a normalization constant.

If for the previous example a PWG representation is used instead as in figure 2(c), the strengths of each of the modes are preserved. This is preferable since the representation would then be consistent with the local statistics determined for each hypothesis.



(a) (b) (c)

Figure 2: (a) shows four recovered modes of a probability distribution together with local statistics. Using a Gaussian sum approximation with components located at the hypotheses would produce the distribution shown in (b), which has only two modes, and also the dominant mode is formed from the cluster of weaker modes. The modes and local variances are however preserved if a piecewise Gaussian approximation is used (c).

While it is possible that a good Gaussian sum approximation may be obtained via a complex fitting process (eg. via the EM algorithm[15]), the PWG representation provides satisfactory approximation at negligible cost of fitting, although sampling from the PWG representation is not as straightforward (discussed later in section 3.3.2).

3.2 Generating Prior Distributions

Obtaining the prior density $p(x_t|Z_{t-1})$ in the next time frame is similar to the Kalman filter prediction step. A dynamical model is applied to the modes of the posterior distribution $p(x_{t-1}|Z_{t-1})$ of the previous time frame to predict the new locations of the modes, followed by increasing the covariances of the Gaussian components according to the process noise. This amount of process noise is dictated by the accuracy of the dynamical model. This may also be viewed as an approximation to the result $p(x_t|Z_{t-1}) = \int_{x_t} p(x_t|x_{t-1})p(x_{t-1}|Z_{t-1})$, where $p(x_t|x_{t-1})$ is a Gaussian centered on the new mode with covariance equal to the process noise covariance. Here x_t is assumed to be conditionally independent of Z_{t-1}.

In the experiments carried out for this paper, we did not use a trained or complex dynamical model. The dynamical model employed is simply a naive constant velocity predictor, and consequently the process noise applied is very high since the prediction is often grossly inaccurate.

3.3 Likelihood Computation

3.3.1 State Probabilities from Image Measurements

In order to model the likelihood $p(z_t|\boldsymbol{x}_t)$, we need to be able to compute the probability that the target figure, when correctly represented by an SPM model with state $\boldsymbol{x}$, generates the image observation z_t in the current frame. This is estimated via

$$p(z_t|\boldsymbol{x}_t) \propto \prod_{\boldsymbol{u}} \exp\left(-\frac{(I(\boldsymbol{u}) - T(\boldsymbol{u}, \boldsymbol{x}_t))^2}{2\sigma^2} \right) \quad (3)$$

where $\boldsymbol{u}$ represent image pixel coordinates, $I(\boldsymbol{u})$ are the image pixel values at $\boldsymbol{u}$, $T(\boldsymbol{u}, \boldsymbol{x})$ are the overlapping template pixel values at $\boldsymbol{u}$ when the SPM model has state $\boldsymbol{x}$, and σ^2 is the pixel noise variance (this has to be known apriori or experimentally obtained). The product is then evaluated for all pixels located within the boundaries of the figure.

Based on (3), it may be observed that the likelihood can be maximized by minimizing $(I(\boldsymbol{u}) - T(\boldsymbol{u}, \boldsymbol{x}_t))^2$. This is achieved through template registration, which may be considered equivalent to recovering the local maximum likelihood solution.

3.3.2 Hypothesis Sampling

We first consider the case of sampling from a single truncated Gaussian. This involves obtaining samples from the original Gaussian distribution, followed by discarding the samples which fall outside the truncation boundary. This may be continued until a satisfactory number of valid samples have been obtained.

The PWG distribution may be equivalently expressed as a union of separate truncated Gaussians with aligned borders, where the borders denote points for which the probability values computed from either Gaussian component on opposite sides of the border are the same (ie. there are no probability discontinuities at the borders). Sampling from the PWG distribution may therefore be carried out with the following steps:

1. Select the ith mode with probability p_i from the set of N modes.
2. Obtain a single sample $\boldsymbol{s}$ from the original Gaussian distribution associated with the ith mode.
3. If $\boldsymbol{s}$ lies within the boundaries of the ith mode (ie. $p(s)$ satisfies (2)), accept the sample; otherwise reject it.
4. Return to step 1 until the required number of accepted samples have been obtained.

3.3.3 State-Space Search for Likelihood Modes

Starting with the initial SPM model states obtained from sampling the prior distribution $p(x_t|Z_{t-1})$, the states are optimized locally in order to converge on the modes of the likelihood $p(z_t|x_t)$. This achieved by maximizing (3), or equivalently by obtaining

$$\arg\min_{\boldsymbol{x}} \; \{ \sum_{\boldsymbol{u}} \left(I(\boldsymbol{u}) - T(\boldsymbol{u}, \boldsymbol{x})\right)^2 \}$$

This is in fact identical to differential template registration of the 2D SPM model whereby the sum of squared pixel residuals is minimized. For this we employ the iterative Gauss-Newton method, which has an advantage of simultaneously recovering the local variances of each mode.

3.4 Deriving Posterior Distributions

Computing the posterior density via (1) involves the multiplication of the prior density $p(\boldsymbol{x}_t|Z_{t-1})$ and likelihood $p(z_t|\boldsymbol{x}_t)$ functions, where both functions are represented in PWG forms as described in the previous sections. The posterior density may be approximated by taking pairs of modes from the prior and likelihood distributions and multiplying the Gaussians independently. This may be further trimmed by selecting only the dominant posterior modes.

To prevent an exponential increase in modes in our experiments, each likelihood mode generates a posterior mode by combining with the most compatible prior mode. This is acceptable as the modes of the likelihood are the dominant factors when a constant velocity predictor with high process noise is used. If a superior predictor is available, greater emphasis may be placed on the prior modes.

4 Experimental Results

The algorithm was tested on three sequences involving Fred Astaire from the movie ‘Shall We Dance’. A 2D 19-DOF SPM model is manually initialized in the first image frame, after which tracking is fully automatic. The augmented state-space in this case has 38 dimensions because the predictor used is a second order auto-regressive (AR) model. Typically the joint probability distribution in the state-space is described via 10 modes in a PWG representation.

In fig. 3, three key frames from an original sequence of eighteen frames are shown, together with the results obtained from using a single mode tracker. Here the stick figure denotes the current state of the tracker. It can be observed that the tracker fails to cope with the ambiguity resulting from self-occlusion when Fred Astaire’s legs cross.

In fig. 4, the multiple modes of the tracker are shown in the top row. The bottom row shows the dominant mode at each frame, which is *solely determined via minimum pixel squared residual error*. This shows the ability of the tracker to handle the ambiguities of self-occlusion by maintaining multiple modes, without even the need for a complex dynamical model.

However, the computational cost of using multiple modes increases at least linearly with the number of modes. In the above case, the single-mode tracker completed the tracking sequence of 18 frames in about 18 seconds. The 10-mode tracker required approximately 2 minutes. Nevertheless the advantage gained from the stability of the tracker is significantly more critical.

5 Previous Work

The first works on articulated 3D tracking were [3, 4]. Yamamoto and Koshikawa [5] were the first to apply modern kinematic models and gradient-based optimization techniques, but their results were limited to 2D motion. Other 3D tracking works include [6, 16, 17, 18]. The work of Ju and et. al. [19] is perhaps the closest to our 2D SPM. Other 2D figure tracking results can be found in [20].

Early applications of Kalman filters (KF) to rigid body tracking appear in [21, 22, 23]. Figure tracking schemes which use the Kalman filter are discussed in [8, 9]. All of these works employ the conventional unimodal KF. One exception is Shimada et. al. [10], in which a simple multiple hypothesis approach is used to handle reflective ambiguity under orthographic projection.

The first applications of classical multiple hypothesis tracking techniques to computer vision problems appeared

Figure 3: Single Mode Tracking Results. Top row: three frames from the original sequence. Bottom row: the single-hypothesis tracker fails to handle the self-occlusion caused by Fred Astaire's legs crossing.

in [24, 25]. An early survey of these techniques can be found in [26]. Recently, Rasmussen and Hager [27] used the joint probabilistic data association filter (JPDAF) [11] to track multi-part objects, such as a face and hand. In contrast to our MHT framework, the JPDAF approach uses a correspondence-based framework for generating hypotheses. Each target is influenced by a linear combination of the resulting measurements.

5.1 Comparisons to Classical MHT and Monte Carlo Methods

Multiple hypothesis tracking was originally developed for radar tracking systems where the measured features are a set of discrete 'blips'. The multiple hypotheses are generated by postulating associations between a single target and each of the different features. In the case of figure tracking there is however no detector for the human figure which explicitly returns features giving different probable skeletal configurations in each image frame. One possible solution would be to consider all combinations of lower-level features, eg. edges obtained from an edge detector, which form high-level 'figure features'. However in scenes with significant clutter, this rapidly leads to an almost intractable number of hypotheses [24, 25]. More importantly, discrete features are not suitable to a large class of problems. For example when using models based on appearance or optic-flow, the data association between the model and image pixels is both probabilistic and continuous – every different set of pixels is a separate feature with a corresponding probability of association to the model. In these instances, classical MHT methods are not applicable.

Instead of using a separate feature-detection process based on image correspondences, our formulation of hypothesis sampling and local state-space search recovers MH states as part of the tracking process. This method is also capable of coping with the above-mentioned problems for which the feature set in continuous. The multiple hypotheses in our method are not simply data-association hypotheses between target and features, but state-space hypotheses which locally maximize the likelihood of the observed image.

Alternatively Monte Carlo methods, such as the CONDENSATION algorithm [12], can be used. These methods express the pdf of the tracker state non-parametrically with a fair set of samples. The number of samples required for accurately modeling the pdf increases with both the dimensionality of the state space and the variance of the pdf, which in the case of tracking is inversely related to the accuracy of the predictor. In our case with 38 state-space dimensions and a weak constant velocity dynamical model, a prohibitive number of samples will be required for reliable tracking with CONDENSATION. A further problem with the sample-based pdf representation is that only the moments of the pdf can be recovered easily. Hence for example while it may be simple to compute the mean state, the maximum likelihood (ML) estimate may not be found accurately, and more significantly the maximum aposteriori (MAP) estimate is difficult to compute.

Experiments carried out using the authors' implementation of the CONDENSATION algorithm bear out these observations. Tracking was attempted on sequences of a person walking using a 26-dimensional tracker based on templates (instead of contours as in [12]). When a *second-order autoregressive (AR) model* trained on walking dynamics was applied, tracking was successful when

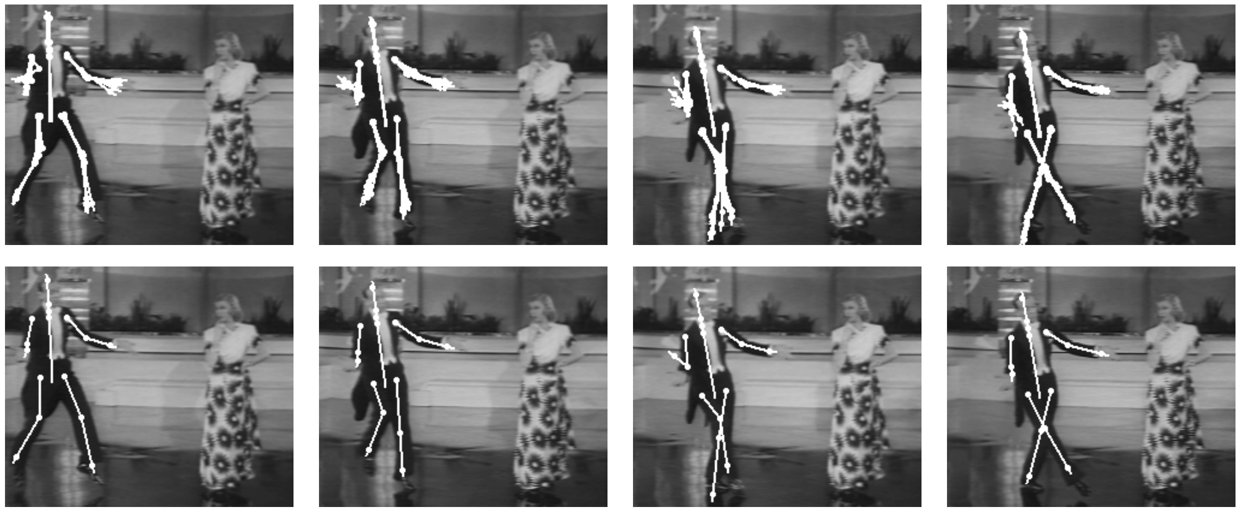


Figure 4: Mode-based Multiple Hypothesis Tracking Results. Top row: the multiple modes of the tracker are shown. Bottom row: the dominant mode is shown, which demonstrate the ability of the tracker to handle ambiguous situations and thus survive the occlusion event.

at least 50 samples were used. However tracking with this AR model can be carried out more efficiently by using our single-hypothesis tracker, with running speeds of 6fps versus 0.4fps. To compare performances when a constant velocity dynamical model was applied instead, we used 200 samples in our CONDENSATION implementation to set the running speed to be approximately equal to our multiple-hypothesis tracker. While the former failed to track after the fourth image frame, our MH tracker was successful for the entire 48 frames.

Our approach copes with weak dynamical models and high-dimensional state spaces by carrying out sample refinement. This allows successful tracking to be achieved with only ten samples. Furthermore because a parametric representation is used throughout the entire process, both the MAP and ML estimates can be recovered easily.

6 Conclusions and Future Work

We have introduced a novel multiple hypothesis tracking algorithm for complex targets with high dimensional state spaces. The key insight is to represent and track the modes in the posterior state density function. These modes are likely to be sparse and separated for visually complex targets such as the human figure. Experimental results from tracking one of Fred Astaire's dance sequences demonstrates the superior performance of our MHT approach over a standard Kalman filter.

In the near future we will present comparative experimental results to that of the CONDENSATION algorithm. We also plan to extend our MHT framework to handle self-occlusions and motion discontinuities in an explicit manner. We will also be investigating the integration of figure tracking with background modeling as well as figure-background segmentation.

References

[1] J. M. Rehg and T. Kanade, "Digiteyes: Vision-based hand tracking for human-computer interaction," in *Proc. of Workshop on Motion of Non-Rigid and Articulated Objects* (J. K. Aggarwal and T. S. Huang, eds.), (Austin, Texas), pp. 16–22, IEEE Computer Society Press, 1994.

[2] D. D. Morris and J. M. Rehg, "Singularity analysis for articulated object tracking," in *Proc. Computer Vision and Pattern Recognition*, (Santa Barbara, CA), pp. 289–296, June 23–25 1998.

[3] J. O'Rourke and N. Badler, "Model-based image analysis of human motion using constraint propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, no. 6, pp. 522–536, 1980.

[4] D. Hogg, "Model-based vision: a program to see a walking person," *Image and Vision Computing*, vol. 1, no. 1, pp. 5–20, 1983.

[5] M. Yamamoto and K. Koshikawa, "Human motion analysis based on a robot arm model," in *Proc. Computer Vision and Pattern Recognition*, pp. 664–665, 1991. Also see Electrotechnical Laboratory Report 90-46.

[6] J. M. Rehg and T. Kanade, "Visual tracking of high dof articulated structures: an application to human

hand tracking," in *Proc. European Conference on Computer Vision*, (Stockholm, Sweden), pp. II: 35–46, 1994.

[7] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Prentice-Hall, 1979.

[8] A. Pentland and B. Horowitz, "Recovery of nonrigid motion and structure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 7, pp. 730–742, 1991.

[9] I. Kakadiaris and D. Metaxas, "Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection," in *Proc. Computer Vision and Pattern Recognition*, (San Francisco, CA), pp. 81–87, June 18–20 1996.

[10] N. Shimada, Y. Shirai, Y. Kuno, and J. Miura, "Hand gesture estimation and model refinement using monocular camera— ambiguity limitation by inequality constraints," in *Proc. International Conference on Automatic Face and Gesture Recognition*, (Nara, Japan), pp. 268–273, April 14–16 1998.

[11] Y. Bar-Shalom and T. E. Fortmann, *Tracking and Data Association*. Academic Press, 1988.

[12] M. Isard and A. Blake, "CONDENSATION – conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[13] T.-J. Cham and J. M. Rehg, "A multiple hypothesis framework for figure tracking," in *Proc. Workshop on Perceptual User Interfaces*, (San Francisco, CA), pp. 19–24, 1998.

[14] T.-J. Cham and J. M. Rehg, "A multiple hypothesis framework for figure tracking," Tech. Rep. CRL 98/8, Compaq Computer Corp. Cambridge Research Lab., Cambridge MA, July 1998.

[15] A. Dempster, N. Laird, and D. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. B39, pp. 1–38, 1977.

[16] J. M. Rehg and T. Kanade, "Model-based tracking of self-occluding articulated objects," in *Proc. of Fifth Intl. Conf. on Computer Vision*, (Boston, MA), pp. 612–617, 1995.

[17] D. M. Gavrila and L. S. Davis, "3-D model-based tracking of humans in action: A multi-view approach," in *Proc. Computer Vision and Pattern Recognition*, (San Francisco, CA), pp. 73–80, June 18–20 1996.

[18] C. Bregler and J. Malik, "Estimating and tracking kinematic chains," in *Proc. Computer Vision and Pattern Recognition*, (Santa Barbara, CA), pp. 8–15, 1998.

[19] S. X. Ju, M. J. Black, and Y. Yacoob, "Cardboard people: A parameterized model of articulated image motion," in *Proc. International Conference on Automatic Face and Gesture Recognition*, (Killington, VT), pp. 38–44, 1996.

[20] Y. Yacoob and L. Davis, "Learned temporal models of image motion," in *Proc. International Conference on Computer Vision*, (Bombay, India), pp. 446–453, January 4–7 1998.

[21] T. Broida and R. Chellappa, "Estimation of object motion parameters from noisy images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, pp. 90–99, 1986.

[22] J. J. Wu, R. E. Wink, T. M. Caelli, and V. G. Gourishankar, "Recovery of the 3-d location and motion of a rigid object through camera image (an Extended Kalman Filter approach)," *International Journal of Computer Vision*, vol. 2, no. 4, pp. 373–394, 1989.

[23] J. L. Crowley, P. Stelmaszyk, T. Skordas, and P. Puget, "Measurement and integration of 3-D structures by tracking edge lines," *International Journal of Computer Vision*, vol. 8, no. 1, pp. 29–52, 1992.

[24] I. J. Cox, J. M. Rehg, and S. Hingorami, "A bayesian multiple hypothesis approach to edge grouping and contour segmentation," *International Journal of Computer Vision*, vol. 11, no. 1, pp. 5–24, 1993.

[25] I. J. Cox and S. L. Hingorani, "An efficient implementation of Reid's Multiple Hypothesis Tracking algorithm and its evaluation for the purpose of visual tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 138–150, February 1996.

[26] B. Rao, "Data association methods for tracking systems," in *Active Vision* (A. Blake and A. Yuille, eds.), ch. 6, pp. 91–105, MIT Press, 1992.

[27] C. Rasmussen and G. D. Hager, "Joint probabilistic techniques for tracking multi-part objects," in *Proc. Computer Vision and Pattern Recognition*, (Santa Barbara CA), pp. 16–21, June 23–25 1998.