COMPAQ

A Multiple Hypothesis Approach to Figure Tracking

Tat-Jen Cham   *James M. Rehg*

Cambridge Research Laboratory

Cambridge Research Laboratory
Technical Report Series

CRL 98/8
July 1998

Cambridge Research Laboratory

The Cambridge Research Laboratory was founded in 1987 to advance the state of the art in both core computing and human-computer interaction, and to use the knowledge so gained to support the Company's corporate objectives. We believe this is best accomplished through interconnected pursuits in technology creation, advanced systems engineering, and business development. We are actively investigating scalable computing; mobile computing; vision-based human and scene sensing; speech interaction; computer-animated synthetic persona; intelligent information appliances; and the capture, coding, storage, indexing, retrieval, decoding, and rendering of multimedia data. We recognize and embrace a technology creation model which is characterized by three major phases:

Freedom: The life blood of the Laboratory comes from the observations and imaginations of our research staff. It is here that challenging research problems are uncovered (through discussions with customers, through interactions with others in the Corporation, through other professional interactions, through reading, and the like) or that new ideas are born. For any such problem or idea, this phase culminates in the nucleation of a project team around a well articulated central research question and the outlining of a research plan.

Focus: Once a team is formed, we aggressively pursue the creation of new technology based on the plan. This may involve direct collaboration with other technical professionals inside and outside the Corporation. This phase culminates in the demonstrable creation of new technology which may take any of a number of forms - a journal article, a technical talk, a working prototype, a patent application, or some combination of these. The research team is typically augmented with other resident professionals—engineering and business development—who work as integral members of the core team to prepare preliminary plans for how best to leverage this new knowledge, either through internal transfer of technology or through other means.

Follow-through: We actively pursue taking the best technologies to the marketplace. For those opportunities which are not immediately transferred internally and where the team has identified a significant opportunity, the business development and engineering staff will lead early-stage commercial development, often in conjunction with members of the research staff. While the value to the Corporation of taking these new ideas to the market is clear, it also has a significant positive impact on our future research work by providing the means to understand intimately the problems and opportunities in the market and to more fully exercise our ideas and concepts in real-world settings.

Throughout this process, communicating our understanding is a critical part of what we do, and participating in the larger technical community—through the publication of refereed journal articles and the presentation of our ideas at conferences–is essential. Our technical report series supports and facilitates broad and early dissemination of our work. We welcome your feedback on its effectiveness.

Robert A. Iannucci, Ph.D.
Director

A Multiple Hypothesis Approach to Figure Tracking

Tat-Jen Cham James M. Rehg

Cambridge Research Laboratory
Compaq Computer Corporation
Cambridge MA 02139

July 1998

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.

Authors email: tjc@crl.dec.com, rehg@crl.dec.com

CRL Technical reports are available on the CRL's web page at
http://www.crl.research.digital.com.

Compaq Computer Corporation
Cambridge Research Laboratory
One Kendall Square, Building 700
Cambridge, Massachusetts 02139 USA

1 Introduction

Visual tracking of human motion is a key technology in a broad range of applications from user-interfaces to video editing. This paper addresses the problem of figure tracking, using a known kinematic model to describe the skeletal constraints [17, 12, 28, 22]. 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 [2] is a classical choice which has been employed in earlier figure tracking work (see [18, 15, 25] 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 multi-modal distributions with Kalman filters [4]. MHT methods have been used with great effectiveness in radar tracking systems, for example. They are designed to process a discrete set of measurements, such as radar returns, at each time instant. A representative approach is Reid's algorithm [24], which employs a bank of Kalman filters to evaluate the combinatoric assignments between discrete measurements and targets. Unfortunately, for visual tracking applications with complex targets the requisite "sensor" typically does not exist. For example, there is no generic figure detector which takes an input image and outputs a set of figure measurements, where each measurement is a different possible skeletal configuration.

An alternative to classical MHT is the class of Monte Carlo methods such as Isard and Blake's CONDENSATION algorithm [13]. These techniques employ a nonparametric sample-based representation of the pdf which can model arbitrary densities. These methods have the disadvantage that the required number of samples grows exponentially with the size of the state space. As a consequence, an accurate dynamic model is required in practice to reduce the number of samples needed for accurate modeling. These factors make nonparametric techniques less attractive for objects like the human figure, which have both a large state space and complex dynamics.

This paper describes a novel MHT formulation which is suitable for figure tracking. The key idea is to explicitly model and track the modes in the state pdf. Our approach is based on the hypothesis that visually complex targets such as the human figure will typically have a small number of well-defined minima in their posterior density. We use a sampling-based state space search process to generate a set of hypotheses corresponding to the local maxima in the likelihood function. By generating hypotheses through state space search we avoid the need for the explicit figure detector required by classical MHT methods. By focusing our probabilistic representation on the modes of the distribution we avoid the explosion in the number of samples that a Monte-Carlo scheme requires. A more detailed comparison can be found in section 5.1. This work is the first application of multiple hypothesis techniques to figure tracking.

2 A Kinematic Model for Figure Registration

Most of the previoous work on articulated object tracking has focused on the use of 3-D kinematic models to estimate the detailed 3-D motion of hands and figures. These approaches require multiple camera viewpoints for accurate estimation and rarely operate on live video (one exception is [22]). In contrast, there are many applications of figure tracking in which only a single camera input is available. One example which motivates this report is the recovery of human motion from movie footage. Another class of examples are vision-based user-interfaces which could benefit from coarse measurements of body pose suitable for gesture recognition, but are unlikely to require accurate 3-D pose recovery.

This report addresses *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 [16]. 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.

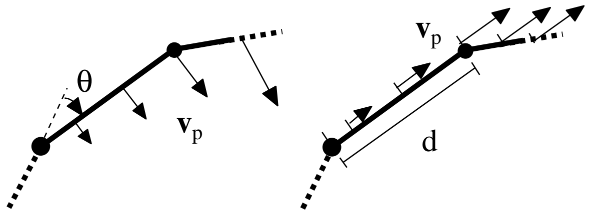


Figure 1: The effect of revolute (θ) and prismatic (d) DOF's on one link from a 2D SPM chain. The arrows show the instantaneous velocity of points along the link due to an instantaneous state change.

A complete discussion of SPM models, including a derivation of the SPM Jacobian and an analysis of its singularities, can be found in [16]. 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.

3 Mode-based Multiple-Hypothesis Tracking

The central goal of a probabilistic tracking framework is to evolve the probability distribution of the tracker state over time. Our approach is based on a parametric representation of the *modes* (local maxima) of the probability density function (pdf) which describes the uncertainty in the state. We use Gaussian kernels to model the pdf in the local neighborhood surrounding each mode and to interpolate between modes. Kernel functions provide a compact description of the pdf, in contrast to the large number of samples a non-parametric method would employ in modeling each mode. Each kernel can be viewed as a hypothesis about the tracker state, establishing a connection with classical MHT methods.

Our adoption of a mode-based representation is based on two assumptions: that the underlying pdf has well-defined modes, and that these modes capture the essential structure of the pdf which is required for accurate tracking. We believe this is a reasonable assumption for complex visual targets like the figure, and the experimental results we present in section 4 support this hypothesis.

The key step in our mode-based tracking algorithm is a technique, called *sample refinement*, for updating the modes of the pdf given an input image. Sample refinement uses sampling from a prior distribution to search for peaks in the likelihood function. It is described in detail in section 3.3.

Our tracking algorithm consists of a series of three steps which are linked through 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-step t are:

1. **Prediction** The prior density $p(x_t|Z_{t-1})$ is generated by passing the modes of $p(x_{t-1}|Z_{t-1})$ through the Kalman filter prediction step.

2. **Likelihood computation** This involves:

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

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

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

3. **Posterior Update** The posterior density $p(x_t|Z_t)$ is computed via Baye's Rule (equation 1) and the set of modes is updated.

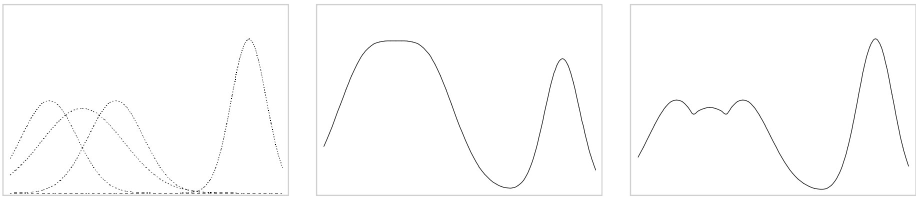
Each of these stages outputs a multimodal description of the state pdf. The piecewise Gaussian representation which we employ in modeling the multimodal pdfs is described in section 3.1. Sections 3.2, 3.3, and 3.4 describe the three stages of our algorithm in detail.

3.1 Piecewise Gaussian Kernels

The use of kernels to model probability density functions is a classical problem in statistics and a wide variety of solutions are available (see [11] for a recent survey). Our current approach is based on the use of Gaussian kernels to model the pdf in the immediate vicinity of each mode. To describe the pdf in the state space regions between the modes we use the *max* function to select the kernel with the highest likelihood. This leads to a piecewise Gaussian representation of the pdf. One advantage of this approach is its computational simplicity, since the kernel parameters are determined entirely by local pdf values. This approach is similar in spirit to locally-weighted regression [3].

We can define the piecewise Gaussian representation as follows: 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 $\boldsymbol{S}_i$.

Given locally fitted Gaussian kernels, one might be tempted to combine them into a Gaussian sum representation by direct superposition. When the modes occur in clumps (which happens frequently) this approach will produce errors, as figure 2 illustrates. A simple example of four hypotheses in a 1-D state-space is shown in figure 2(a). If the hypotheses are summed the combined pdf has only two modes, as shown in figure 2(b). This results in a cluster of weaker modes being over-represented at the expense of strong but isolated modes.



(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).

We employ 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.

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

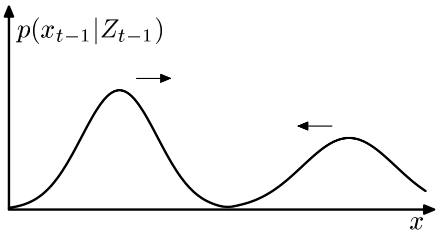
where k is a normalization constant.

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

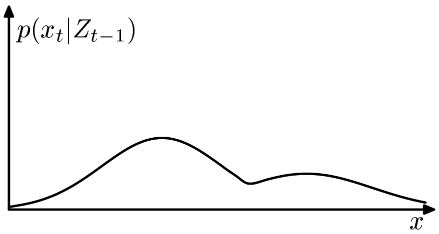
An accurate Gaussian sum representation could be obtained through a more complex and costly fitting process using the EM algorithm [21]. However, we have found that the PWG representation provides satisfactory approximation at a greatly reduced cost of fitting. This representation does have two disadvantages: Sampling from the PWG representation and propagating it through a dynamic model are not as straightforward as they would be for a Gaussian mixture model. These points are discussed further in sections 3.2 and 3.3.1.

3.2 Prediction

The prior density $p(x_t|Z_{t-1})$ in the next time frame is obtained by applying the Kalman filter prediction step to each of the modes of the posterior distribution $p(x_{t-1}|Z_{t-1})$ in the previous time frame. A dynamical model predicts the new locations of the modes, while the covariances of the Gaussian components are increased according to the process noise. The amount of process noise is determined by the accuracy of the dynamical model. This process is illustrated in Figures 3(a) and (b), which show a 1-D distribution before and after prediction. This formulation 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.



(a)



(b)

Figure 3: Prediction step of MHT algorithm. Two modes of a 1-D density for x (a) are extrapolated to the next time step (b).

A disadvantage of using the PWG representation is that the application of the standard Kalman filter steps to individual modes for computing prior and posterior distributions is only mathematically correct for a Gaussian sum parameterization. However, the Kalman filter steps are reasonable approximations in a PWG representation when the significant modes of the distribution are well-defined with small local variances. This is the situation encountered when observation noise is low and the hypotheses represent discrete well-defined ambiguous configurations, as opposed to the situation when observation noise is high where ambiguous configurations are fused and continuous in nature. In the sequences used for testing our tracker, the separate hypotheses result from clutter and self-occlusions rather than camera noise. This justifies the use of the PWG representation within the tracking framework.

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

At the heart of any visual tracking problem is the computation of the likelihood of the observed data given a state model. Data likelihood is the fundamental source of multimodality in visual tracking, as there are typically many different state configurations that are consistent with a given set of image measurements. Our approach is to describe the likelihood surface using a collection of Gaussian kernels. This section describes the algorithm for computing kernel positions and parameters given a predicted state pdf and an input image.

3.3.1 Hypothesis Sampling

We employ sampling from the predicted state pdf to generate starting points for the local search process that identifies the modes in our likelihood representation. We first consider the case of sampling from a single truncated Gaussian. This involves obtaining samples from the original Gaussian distribution (eg. we used publicly available code based on [1]), 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. Figure 4(a) shows a representative outcome of the sampling process for the predicted distribution shown in Figure 3(b).

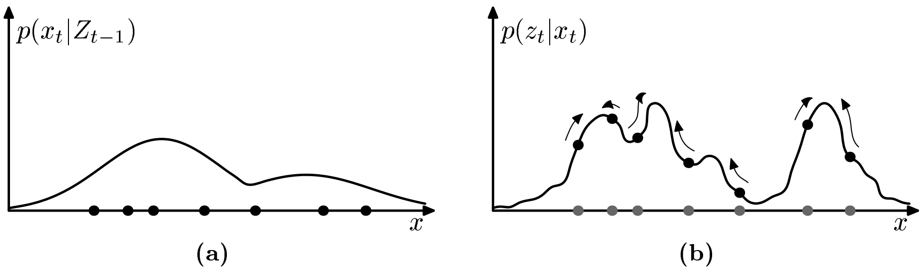


Figure 4: (a) Samples drawn from the 1-D density of Figure 4 are shown as dots. (b) Each sample seeds a local search of the likelihood function.

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 (using notation defined in section 3.1).
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.

3.3.2 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

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

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 associated with the modes. This search process is illustrated in Figure 4(b). Arrows show the direction of steepest ascent from each seed point. Note that a given model may attract multiple seed points.

3.3.3 State Probabilities from Image Measurements

Given the detected modes of the likelihood surface, the final step in computing the likelihood model is the estimation of the parameters of the Gaussian modes. This can be accomplished using an image noise model which gives 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. The model can be written

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

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.

The final PWG representatation of the likelihood is illustrated in Figure 5 for the example of Figure 4. The detected modes are shown as black circles. The PWG surface overlays the "true" likelihood surface, and the seed points are drawn in gray. Approximation error is greatest where the modes are close together.

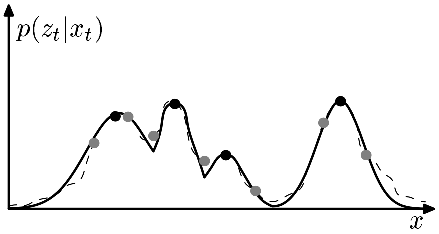


Figure 5: Piece-wise Gaussian likelihood model constructed from the detected mode positions shown as black circles. The true likelihood (from Figure 4) is shown as a dashed line.

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.

3.4 Posterior Update

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.

However in our experiments, the posterior density is taken to be identical to the likelihood. This simplification is acceptable because we used a simple constant velocity predictor with correspondingly high process noise. The modes of the likelihood are the dominant factors in this case. If a superior predictor were available, better results could be obtained by modeling the posterior density more accurately.

3.4.1 Posterior versus Likelihood Distributions

An important point to note is while the posterior density incorporates all available information at the end of each time frame, it may also be useful to retain the likelihood distributions as well. This is true when for example an off-line process is available to refine the tracking using further knowledge not available to the original tracker. This refinement may be achieved with more advanced target and dynamical models as well as using the observations in batch mode rather than in a sequential manner. For example in the figure tracking scenario, a 3D kinematic model with angular and length constraints may be employed off-line to improve on the initial tracking made with a 2D SPM model; additionally more accurate 2D or 3D dynamical models may be used to improve the tracking made using a simple constant velocity model. The original posterior distributions should not be used as input since they incorporate erroneous prior knowledge which is superseded when the improved models are used. Instead, the likelihood distributions can be used as input because they encode solely the information obtained from observations made within each time frame.

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. 6, 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. 7, 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.

Figure 6: 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.

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.

Figure 7: 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.

5 Previous Work

The first works on articulated 3D tracking were [17, 12]. Yamamoto and Koshikawa [28] 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 [22, 23, 10, 5]. The work of Ju and et. al. [14] is perhaps the closest to our 2D SPM. Other 2D figure tracking results can be found in [27].

Early applications of Kalman filters (KF) to rigid body tracking appear in [6, 26, 9]. Figure tracking schemes which use the Kalman filter are discussed in [18, 15]. All of these works employ the conventional unimodal KF. One exception is Shimada et. al. [25], 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 in [8, 7]. An early survey of these techniques can be found in [19]. Recently, Rasmussen and Hager [20] used the joint probabilistic data association filter (JPDAF) [4] 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 [8, 7]. 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 [13], 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 predictor, 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.

Our approach copes with weak predictors 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.H. Ahrens and U. Dieter. Extensions of Forsythe's method for random sampling from the normal distribution. *Mathematical Computing*, 27(124):927–937, 1973.

[2] Brian D. O. Anderson and John B. Moore. *Optimal Filtering*. Prentice-Hall, 1979.

[3] Christopher G. Atkeson, Andrew W. Moore, and Stefan Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11:11–73, 1997.

[4] Yaakov Bar-Shalom and Thomas E. Fortmann. *Tracking and Data Association*. Academic Press, 1988.

[5] Christoph Bregler and Jitendra Malik. Estimating and tracking kinematic chains. In *Proc. Computer Vision and Pattern Recognition*, pages 8–15, Santa Barbara, CA, 1998.

[6] Ted Broida and Rama Chellappa. Estimation of object motion parameters from noisy images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:90–99, 1986.

[7] Ingemar J. Cox and Sunita 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*, 18(2):138–150, February 1996.

[8] Ingemar J. Cox, James M. Rehg, and Sunita Hingorami. A bayesian multiple hypothesis approach to edge grouping and contour segmentation. *International Journal of Computer Vision*, 11(1):5–24, 1993.

[9] James L. Crowley, Patrick Stelmaszyk, Thomas Skordas, and Pierre Puget. Measurement and integration of 3-D structures by tracking edge lines. *International Journal of Computer Vision*, 8(1):29–52, 1992.

[10] Dariu M. Gavrila and Larry S. Davis. 3-D model-based tracking of humans in action: A multi-view approach. In *Proc. Computer Vision and Pattern Recognition*, pages 73–80, San Fransisco, CA, June 18-20 1996.

[11] Neil Gershenfeld. *The Nature of Mathematical Modeling*. Cambridge University Press, 1999. To appear.

[12] David Hogg. Model-based vision: a program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.

[13] Michael Isard and Andrew Blake. CONDENSATION – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[14] Shannon X. Ju, Michael J. Black, and Yaser Yacoob. Cardboard people: A parameterized model of articulated image motion. In *Proc. International Conference on Automatic Face and Gesture Recognition*, pages 38–44, Killington, VT, 1996.

[15] Ioannis A. Kakadiaris and Dimitris Metaxas. Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. In *Proc. Computer Vision and Pattern Recognition*, pages 81–87, San Fransisco, CA, June 18-20 1996.

[16] Daniel D. Morris and James M. Rehg. Singularity analysis for articulated object tracking. In *Proc. Computer Vision and Pattern Recognition*, pages 289–296, Santa Barbara, CA, June 23–25 1998.

[17] J. O'Rourke and N. Badler. Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(6):522–536, 1980.

[18] Alex Pentland and Bradley Horowitz. Recovery of nonrigid motion and structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):730–742, 1991.

[19] Bobby Rao. Data association methods for tracking systems. In Andrew Blake and Alan Yuille, editors, *Active Vision*, chapter 6, pages 91–105. MIT Press, 1992.

[20] Christopher Rasmussen and Gregory D. Hager. Joint probabilistic techniques for tracking multi-part objects. In *Proc. Computer Vision and Pattern Recognition*, pages 16–21, Santa Barbara CA, June 23–25 1998.

[21] R. Redner and H. Walker. Mixture densities, maximum likelihood, and the EM algorithm. *SIAM Revew*, 26:195–239, 1994.

[22] James M. Rehg and Takeo Kanade. Visual tracking of high dof articulated structures: An application to human hand tracking. In Jan-Olof Eklundh, editor, *Proc. European Conference on Computer Vision*, volume 2, pages II: 35–46, Stockholm, Sweden, 1994. Springer-Verlag.

[23] James M. Rehg and Takeo Kanade. Model-based tracking of self-occluding articulated objects. In *Proc. of Fifth Intl. Conf. on Computer Vision*, pages 612–617, Cambridge, MA, 1995.

[24] D. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 24(6):843–854, 1979.

[25] Nobutaka Shimada, Yoshiaki Shirai, Yoshinori Kuno, and Jun 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*, pages 268–273, Nara, Japan, April 14–16 1998.

[26] 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*, 2(4):373–394, 1989.

[27] Yaser Yacoob and Larry Davis. Learned temporal models of image motion. In *Proc. International Conference on Computer Vision*, pages 446–453, Bombay, India, January 4–7 1998.

[28] Masanobu Yamamoto and Kazutada Koshikawa. Human motion analysis based on a robot arm model. In *Proc. Computer Vision and Pattern Recognition*, pages 664–665, 1991. Also see Electrotechnical Laboratory Report 90-46.

A Multiple Hypothesis Approach to Figure Tracking

Tat-Jen Cham James M. Rehg

CRL 98/8
July 1998