Joint Probabilistic Techniques
for Tracking Multi-Part Objects

Christopher Rasmussen   Gregory D. Hager

Center for Computational Vision & Control
Yale University
New Haven, CT 06520-8267

Abstract

Common objects such as people and cars comprise many visual parts and attributes, yet image-based tracking algorithms are often keyed to only one of a target's identifying characteristics. In this paper, we present a framework for combining and sharing information among several state estimation processes operating on the same underlying visual object. Well-known techniques for joint probabilistic data association are adapted to yield increased robustness when multiple trackers attuned to disparate visual cues are deployed simultaneously. We also formulate a measure of tracker confidence, based on distinctiveness and occlusion probability, which permits the deactivation of trackers before erroneous state estimates adversely affect the ensemble. We will discuss experiments focusing on color-region- and snake-based tracking that demonstrate the efficacy of this approach.

1   Introduction

More powerful computing hardware and new vision algorithms have expanded the scope of tracking research from its origins in simple geometric shapes to include such complex objects as people and automobiles. For many tasks, techniques for tracking generic edges, curves, blobs, and textures have proven to be applicable with minor modifications to tracking hands, arms, heads, faces, and cars [2, 4, 5, 7].

Despite these advances, most visual tracking algorithms are quite brittle. In particular, many systems are easily confused in commonly occurring visual situations because of their reliance on a single cue or methodology for locating their target. Consider the problem of tracking a person with the goal of providing not only a rough guess of where they are, but also of furnishing information about the current posture of the head, torso, limbs, and so forth. The articulation of human bodies makes self-occlusion [9] (where one part of the body moves in front of the other) and self-distraction (when similar parts—e.g., the hands—are close to one another) common challenges to robust state estimation. Moreover, in many situations other moving objects and variegated backgrounds can further aggravate problems of occlusion and distraction [1, 7]. As recent work in multi-cue tracking suggests [12], one way toward robust visual tracking is through exploiting several simultaneously measured visual cues in as flexible a fashion as possible.

For example, a person tracker that regards its target as consisting of two colored regions—a flesh-colored face above a red-colored shirt—and a head silhouette, represented by a snake. The tracker may rely heavily on the red shirt to maintain contact when the person is surrounded by other, distracting faces in a crowded room. Using *a priori* knowledge of the geometric relationship between a standing person's torso and head, a rough fix on the image position of the head can be derived from the shirt's image location and scale. If the person walks behind a piece of furniture, leaving only their face visible, the tracker can switch its focus to this part of their body. When the person walks in front of a highly-textured background, the snake may become confused, increasing the tracker's reliance on color cues. If the background is a tan brick wall similar in color to skin, the edge cues used by the snake will be sufficient for disambiguation.

In short, attending to multiple cues associated with an object can alleviate many difficulties. Approaches to tracking in this spirit have been successful [3, 13], but as yet little work has been done toward creating an extensible system for tracking increasingly complex, multi-part objects through a wide range of poses, backgrounds, and lighting conditions. In this paper we develop a framework for constructing vision-based tracking systems that rely on multiple visual cues and part-based decompositions to track complex objects.

The probabilistic and joint probabilistic data association filters introduced in [1] serve as a starting point for developing multi-part, multi-attribute tracking methods. We show how object state estimation using a mixture of color region and snake trackers [8, 11] can be made less sensitive to distraction (clutter) by exploiting inter-part relationships, and also how target occlusion can be accommodated through measures for deciding to "switch" a component tracking algorithm on or off, which we term *variable tracker activation*.

2 Data Association Filters

The probabilistic data association filter (PDAF) [1] is an extension of the Kalman filter [1] that casts the problem of data association, or how to update the state when there are multiple measurements and a single target, in a Bayesian framework. One step in the Kalman filter is the computation of the innovation $\nu = \mathbf{z} - \hat{\mathbf{z}}$, where $\mathbf{z}$ is the observed measurement and $\hat{\mathbf{z}}$ is the one predicted from the current state $\mathbf{X}$ by the measurement equation $\hat{\mathbf{z}} = \mathbf{H}\,\mathbf{X}$ [1]. The PDAF introduces the notion of the *combined* innovation, computed over the n measurements detected at a given time step as the weighted sum of the individual innovations: $\nu = \sum_{i=1}^{n} \beta_i \nu_i$. Each β_i is the probability of the association *event* θ_i that the ith measurement is target-originated. Also computed is β_0, the probability of the event that none of the measurements is target-originated. These events encompass all possible interpretations of the data, so $\sum_{i=0}^{n} \beta_i = 1$. The association probabilities β_i are derived from a uniform noise model for spurious measurements and an assumed normal PDF on the correct measurement. Details are given in [1].

The PDAF also develops the idea of a *validation gate*, or an ellipsoidal volume in measurement space, derived from the current estimate and uncertainty of the target state, such that the probability of a target-originated measurement appearing outside of it is negligible. Little accuracy is thus lost by disregarding measurements falling outside the gate. Using a tracking window to limit target search is a common approximation of the validation gate

2.1 Joint PDAF

The distractor model used by the PDAF to calculate each association probability β_i assumes that the target-originated measurement is the only persistent one in the environment. This is a questionable assumption for many distractors, but it certainly does not hold for multi-part objects. Because of the spatial proximity of the parts, one target-originated measurement may often fall within another target's over-

lapping validation gate. Such persistent interference, were one to simply run a separate PDAF on each part, could lead to multiple trackers locked onto the same part.

The joint probabilistic data association filter (JPDAF) [1] deals with this problem by sharing information among separate PDAF trackers in order to more accurately calculate association probabilities. The essential result is an exclusion principle of sorts that prevents two trackers from latching onto the same target.

A key notion in the JPDAF is of a *joint event* Θ, or conjunction of possible target-measurement pairings Θ_{jt_j}, where t_j is the index of the target to which measurement j is matched. Because the expression of joint event probabilities is simplified by using the entire surveillance region as each target's validation gate, efficiency is achieved by considering only *feasible* joint events. The two criteria for a feasible joint event are that each measurement has exactly one source and that the number of measurements associated with each target t is either 0 or 1. Accordingly, we define τ_j to be 0 if measurement j is attributed to noise and 1 if it is associated with a target.

Let $\omega_{jt}(\Theta) = 1$ if $\Theta_{jt} \subset \Theta$ and 0 otherwise. Then the probability of association between measurement j and target t given measurements Z is given by $\beta_{jt} = \sum_{\Theta} P(\Theta\,|\,Z)\omega_{jt}(\Theta)$, where:

$$P(\Theta\,|\,Z) = \kappa \prod_{j=1}^{n} [N_j]^{\tau_j} \prod_{t=1}^{T} \gamma_t. \qquad (1)$$

κ contains terms for normalization and scaling, γ_t is a prior probability on target t being visible (see [1] for details), and N_j is the Gaussian PDF $N[\mathbf{z}_j; \hat{\mathbf{z}}^{t_j}, \mathbf{S}^{t_j}]$ for measurement j ($\mathbf{z}_j$ is the measurement value, $\hat{\mathbf{z}}^{t_j}$ is the predicted measurement value for target t_j, and $\mathbf{S}^{t_j}$ is the associated innovation covariance). State estimation is then the same as for the PDAF.

3 Constrained JPDAF: Parts

We define a *part* as a spatially distinct sub-target physically attached to the object of interest—e.g., hands and a face are parts of the human body. The JPDAF, originally developed to track aircraft radar returns, does not provide for any constraints on targets to maintain a particular configuration. Such a stipulation could help to distinguish a complex tracked object from the background or other objects. This capability is added by altering the calculation of the probability of a joint event given in (1) to also quantify how well the measurements fit a multi-part object model.

Intuitively, an object model describes how the likelihood of one part of an object being in a certain state depends on the states of the other parts. Suppose we let z_t be the index of the measurement associated with target t. A model for an object comprising T parts p_i can be embedded within a probability function $C(\mathbf{Z}, \mathbf{X})$ that quantifies the degree of fit over a given set of feasible matches between the object parts' states $\mathbf{X} = \{\mathbf{X}^t\}$ and the measurements $\mathbf{Z} = \{\mathbf{z}_{z_t}\}$ matched to them. Here we consider the case where C can be decomposed into a product of pairwise constraint probability functions $C_{ij}(\mathbf{z}_{z_i}, \mathbf{z}_{z_j}, \mathbf{X}^i, \mathbf{X}^j)$ (denoted C_{ij}) such that $C(\mathbf{Z}, \mathbf{X}) = \prod_{i=1}^{T} \prod_{j=1}^{T} C_{ij}$. The absence of a constraint between two parts p_i and p_j is indicated by $C_{ij} = 1$. We let $C_{ii} = 1$, and allow $C_{ij} \neq C_{ji}$.

We insert inter-part constraints into the filter equations by modifying (1) to become:

$$P(\Theta \mid Z) = \kappa \prod_{j=1}^{n} \left[N_j \prod_{i=1}^{n} [C_{t_i t_j}]^{\tau_i} \right]^{\tau_j} \prod_{t=1}^{T} \gamma_t. \quad (2)$$

For each measurement $\mathbf{z}_j$, the product containing $C_{t_i t_j}$ cycles through every other measurement $\mathbf{z}_i$, accumulating how well the relationship between them matches the constraint between their associated targets.

As an example constraint function C, consider an object composed of n rigidly linked parts, restricted to translations parallel to the image plane. If measurements for all part trackers are simply image coordinate pairs, then the physical constraints of the system can be captured by a set of image vectors between parts. For each part pair p_i, p_j, an expected measurement difference vector $\mu_{ij} = \hat{\mathbf{z}}_0^j - \hat{\mathbf{z}}_0^i$ is computed from some canonical state $\mathbf{X}_0$, as well a covariance Σ_{ij} on the expected measurement difference. Then we can define a Gaussian $C_{ij}(\mathbf{z}_{z_i}, \mathbf{z}_{z_j}, \mathbf{X}^i, \mathbf{X}^j) = N[\mathbf{z}_{z_j} - \mathbf{z}_{z_i}; \mu_{ij}, \Sigma_{ij}]$.

Since the case above is a rigid linkage, one could use a single position vector in $\Re^2$ to describe the system [10, 11]. However, we have found that our formulation, which simply biases a probabilistic state estimator to favor an interpretation of the data that best matches the target model, works quite well while retaining a useful degree of modularity and flexibility. We note that it is always possible to combine the information contained within the separate state vectors to obtain a single, consolidated state estimate if desired.

3.1 Color Regions as Parts

A uniformly colored region [8] part is formally defined by pixel membership in a five-dimensional ellipsoid in image-RGB space with center μ and scale and
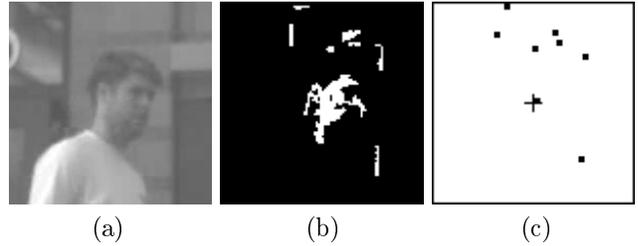

(a)   (b)   (c)

Figure 1: Color Regions and PDAF (from an MPEG). (a) Tracking window; (b) Largest connected components of flesh color; (c) Measurements derived from their centroids.

rotation given by Σ. For reasons explained in [8], the state $\mathbf{X}$ of a color part is restricted to the ellipsoid center $\mu = [x, y, r, g, b]^T$, while Σ is retained as a fixed parameter. The state is initialized by computing the principal components of manually-sampled target pixels. We have found that a stationary dynamical model with relatively high process noise often works well for tracking people's body parts.

For the filter update at time t, measurements are first validated by eliminating image pixels outside the ellipsoid $[\mu_t, \Sigma]$ (a rectangular tracking window serves as the image-spatial gate). To facilitate computation of the association probabilities β, the remaining pixels must be converted to point-like measurements. Each pixel could be a separate measurement, but this would be combinatorially cumbersome and it loses the concept of a region. Instead, the mean positions and colors of the largest connected components (CC) of the validated pixels are used as measurements. This approximation gives good results as long as each CC is relatively compact. The process is illustrated in Figure 1.

Application of any of the above data association filters is straightforward after the completion of these steps. Our implementation of a constrained JPDAF tracker uses the same C_{ij} as the example in the previous section, except that measurements have an additional $[r, g, b]^T$ color component, increasing the dimensionality of the Gaussian. This constraint model is an adequate description of the situation when tracking a person's face, shirt, and pants while sitting or walking. Just one rigid constraint is often sufficient to discriminate an object in an otherwise distracting situation. Parts attached in a non-rigid way, such as hands, can be incorporated by specifying only a weak proximity constraint between them and other parts.

Figure 2 shows the utility of the JPDAF and constrained JPDAF for avoiding mistracking. Color-based trackers are initialized on a person's hand, face, and shirt; the hand then passes in front of the face and
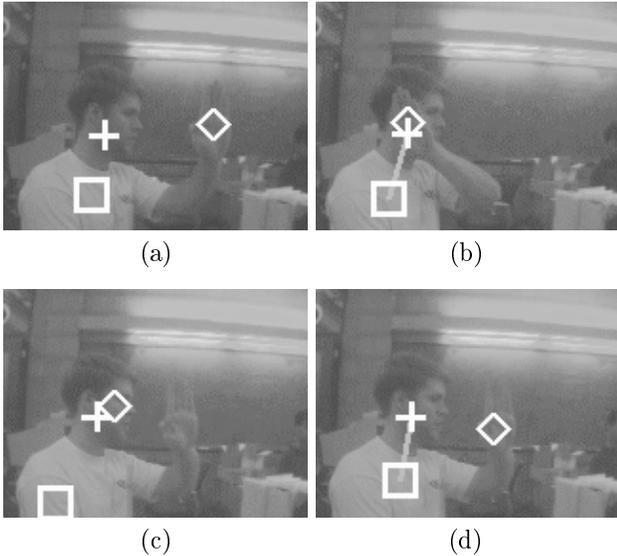
Figure 2: Avoiding distraction (from an MPEG). (a) Initial states of color parts; (b) Face and hand overlap (from constrained JPDAF sequence); (c) PDAF: hand tracker sticks to face; (d) JPDAF with constraint between face and shirt: trackers stay with correct targets.

moves away. When running as independent PDAF's the hand tracker, attracted to skin color, often "sticks" to the face after the hand is removed, and vice versa. Using the same image sequence, we confirmed that a JPDAF avoids this problem because the possibility of the hand and face tracker locking onto the same color region is excluded as an infeasible joint event. However, the JPDAF does not prevent the hand and face trackers from switching places when their paths cross. A constraint to prefer candidate face regions at a vertical offset from the current shirt tracker state effectively anchors the face tracker to the shirt. Although the hand cannot be distinguished from the face while they are overlapping, when it is moved away from the expected face position it is disfavored.

4 Heterogeneous Cues: Attributes

Thus far we have limited our discussion of tracking objects using multiple cues to collections of spatially distinct instances of the same method. We now discuss another kind of cue, which we call an *attribute*, for which multiple trackers "clumping" on the same target may actually be desirable. An attribute is a visual characteristic such as color, edges, texture, depth, or motion. Fundamentally, a part is *what* a tracker tracks, while an attribute is *how* the tracker identifies its target. By its nature, a single part can possess multiple attributes, so it does not make sense to re-

tain a JPDAF-style exclusion principle that prevents multiple trackers of different modalities from following the same target. However, constraints do apply: a color region tracker and a B-spline snake [2] both locked onto a hand, for instance, could be expected to have coincident centers of image mass, or the angle of the major axis of the region could be expected to agree with that of the B-spline.

Different kinds of trackers have distinct measurement spaces, so a separate JPDAF tracker is run for each attribute. Nonetheless, constraint information must be communicated between each group of same-attribute trackers. Suppose we have an object consisting of T_a parts for each attribute a out of m total attributes. Then let C_{ij}^{ab} be the constraint function between the ith part of the ath attribute and the jth part of the bth attribute. It follows that $C_{ij}^{aa} = C_{ij}$, the familiar single JPDAF inter-part constraint. If n_a measurements are detected for the ath attribute, we can modify (2) as follows for the constrained JPDAF on that attribute:

$$P_a(\Theta \mid Z) = \kappa \prod_{j=1}^{n_a} \left[N_j \prod_{b=1}^{m} \prod_{i=1}^{n_b} [C_{t_i^a t_j^b}^{ab}]^{\tau_i^b} \right]^{\tau_j^a} \prod_{t=1}^{T_a} \gamma_t. \quad (3)$$

The superscripts on the τ's, t_i and tj are to clarify which attribute generated the measurement or to which set of parts the target belongs, since there are m different sets of measurements. Any other variables that implicitly refer to a particular attribute should be assumed to use a. Note that this formula reduces to (2) when there is only one attribute ($m = 1$).

4.1 Snakes as Parts

An attribute well-suited to combination with color regions is snake tracking [2, 11]. We follow [2] by representing a snake as an affine parametrization of a B-spline (only translation is used for state in the experiment below). Linking edge fragments into contours to derive a set of discrete measurements in a manner similar to the color pixel grouping by connectivity is combinatorially problematic. Rather, we employ an approach adapted from the state sampling technique used in the Condensation algorithm [5]. First, n minimally separated samples ($n \geq 100$) are generated from a normal distribution in the target snake's state space, centered on its current state $\mathbf{X}$. The value of n and the variance along the translational (image) axes are chosen to give adequate image coverage to a circular "tracking window" about the snake. Each sample snake is scored based on how well it corresponds to an actual image contour at that location (using $p(z \mid x)$ as
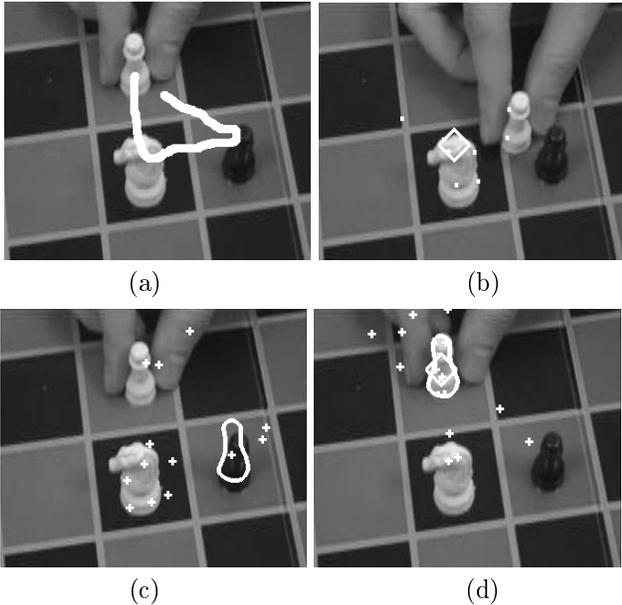
Figure 3: Multi-attribute tracking (from an MPEG). (a) The white line traces the true path of a tracked chess pawn; (b) A color PDAF tracker (whose state is indicated by the $\Diamond$) is distracted by the white knight; (c) A snake PDAF tracker (whose contour is highlighted) is distracted by the black pawn; (d) A joint color region and snake tracker overcomes both distractions. Color and snake measurements are overlaid as boxes and crosses, respectively.

described in [5]). The top k scoring samples ($k \ll n$) are projected to measurement space as putative snake locations. Minimal image separation between the samples is enforced in order to reduce the phenomenon of multiple samples being drawn from the same underlying image feature.

Figure 3 illustrates the advantage of utilizing multiple attributes. Tracking the white pawn as it moves behind the white knight and the black pawn is a difficult task for PDAF trackers relying solely on color or shape as cues. Such trackers can be distracted by nearby image features similar in that attribute, as Figures 3(b) and (c) show. Constraining a snake and color region tracker to prefer coincident centers using (3), however, enables both to successfully follow the target along its entire course, as shown in Figure 3(d).

5   Variable Tracker Activation

Tracking *failure* [12] occurs when contact with a target is lost, either from occlusion or because clutter distracts the tracker away from the true target. The JPDAF and constrained JPDAF try to prevent failure due to distraction, but they can not completely eliminate it. When tracking an object with multiple parts

or attributes, the utility of each cue relative to current image conditions is thus used to weight the various methods to avoid bias. By ceasing state estimation temporarily for any failing parts, erroneous state estimates may be prevented from propagating from failing trackers to healthy ones linked by constraints.

Our framework for recognizing and managing partial failure follows from a notion of tracker *confidence*, or a tracker's self-estimated probability of mistracking based on image conditions and its capabilities. Many phenomena bear on confidence, but here we limit ourselves to two parameters that can be estimated heuristically fairly easily. We quantify the confidence χ of the tracker of a part p_i as a combination of the estimated probabilities of part occlusion $P(O_{p_i})$ and distraction $P(D_{p_i})$, the latter of which can be characterized as a function of part distinctiveness. Confidence is only as high as the greatest source of uncertainty allows, so $\chi(p_i) = 1 - \max(P(O_{p_i}), P(D_{p_i}))$. Details of the calculation of $P(O_{p_i})$ and $P(D_{p_i})$ under the constrained JPDAF framework are given below.

When $\chi(p_i)$ of the tracker for a part p_i with constraint links at least one other part falls below some threshold $\rho \in [0,1]$, it is *deactivated*. This means that its image-based state estimation is switched off discretely, and $C_{ij} = 1$ for purposes of calculating (2). The state of p_i is instead chosen to maximize $\prod_{j=1}^{T} C(\hat{\mathbf{z}}^j, \hat{\mathbf{z}}^i, \mathbf{X}^j, \mathbf{X}^i)$. For the rigid object described in Section 4, this works out to the average of the states of those parts constraining p_i plus their respective state difference vectors. While a tracker is deactivated it continues to perform confidence estimation until $\chi(p_i) \geq \rho$, at which point it is *reactivated*.

The probability of a target t being occluded can be derived directly from the JPDA filter as $P(O_t) = \beta_{0t}$, the probability that none of the observed measurements are associated with the target. We estimate the probability of a tracker being distracted by measuring the distinctiveness of its target in the image environment. With the JPDAF, a target t's distinctiveness depends on how much more probable one association is than the rest. We are only interested in positive associations, or those for which $j > 0$, because if $\beta_{0t} = 1$ then the target is not visible, which is equivalent to saying that it cannot be distinguished from the background. Therefore, if $\beta_{0t} = 1$ then $P(D_t) = 0$.

Otherwise, we normalize each of the remaining β_{jt}'s so that they sum to 1, yielding $\hat{\beta}_{jt}$. The peakedness of these modified association probabilities $\hat{\beta}_{jt}$ indicates the degree of certainty about the best match choice, and can be captured to a large extent by the entropy:
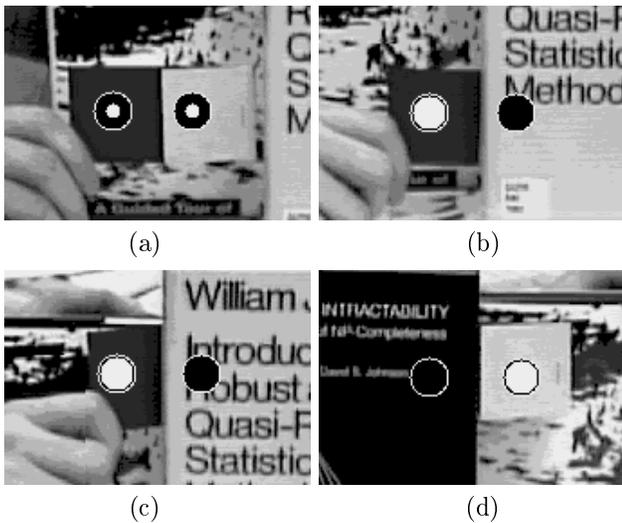
Figure 4: Constrained JPDAF color region tracker exhibiting activation/deactivation due to occlusion. (a) Parts p_1, p_2 unoccluded, active; (b) Part p_2 occluded, inactive; (c) Inactive part p_2's state passively estimated from active part p_1 as object moves; (d) Part p_1 becomes occluded, inactive as part p_2 is reactivated.

$$H_t = -\sum_{j=1}^{n} \hat{\beta}_{jt} \log \hat{\beta}_{jt} \qquad (4)$$

When $\hat{\beta}_{jt} = 1$ for some j, entropy is minimized because there is no uncertainty about which association to make. When every $\hat{\beta}$ is equal, the entropy is maximal, meaning that no association is more likely than any other. We convert the entropy of a target t to a probability of distraction with the formula $P(D_t) = H_t/H_{max}$, where $H_{max} = -\log \frac{1}{n}$ is the maximal entropy for the current number of measurements.

The tracker activation framework is demonstrated in Figure 4 for a two-part object with a rigid constraint. Despite occlusions first of one part and then the other, tracking proceeds smoothly through multiple activation switches as the state of the currently occluded part is derived from its unoccluded partner or the image as conditions warrant. A white inside black circle indicates normal state estimation, a black circle indicates passive estimation via a constraint, and a white circle indicates normal state estimation while driving another part's passive estimation.

6 Conclusion

We believe the approach described in the paper is a promising step toward developing visual tracking systems which flexibly combine multiple parts and attributes to track complex targets in unstructured environments. Our current work is proceeding in a variety of directions. Although phrased in terms of classical linear state-estimation techniques, many of the concepts carry over to other types of state estimation—e.g., the Condensation algorithm [5, 6]. Experimentally, we are adapting the Condensation algorithm to perform joint color region and contour tracking with the goal of comparing its computational and statistical performance vs. modified JPDAF. Also, the image-to-measurement-set methods for color regions and snakes have thus far been largely heuristically motivated. We hope to base these criteria on a more fundamental problem statement. Finally, we are working to extend our techniques to further tracking modalities and more complex inter-part constraint relationships.

7 Acknowledgments

References

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

[2] A. Blake, M. Isard, and D. Reynard. Learning to track the visual motion of contours. *Artificial Intelligence*, No. 78, pp. 101-133, 1995.

[3] C. Bregler. Learning and Recognizing Human Dynamics in Video Sequences. In *CVPR '97*, pp. 568-574, 1997.

[4] G. Hager and P. Belhumeur. Real-Time Tracking of Image Regions with Changes in Geometry and Illumination. In *CVPR '96*, pp. 403-410, 1996.

[5] M. Isard and A. Blake. Contour Tracking by Stochastic Propagation of Conditional Density. In *ECCV '96*, pp. 343-356, 1996.

[6] M. Isard and A. Blake. A Mixed-state Condensation Tracker with Automatic Model-switching. To appear in *ICCV '98*, 1998.

[7] D. Koller, J. Weber, and J. Malik. Robust Multiple Car Tracking with Occlusion Reasoning. In *ECCV '94*, pp. 189-196, 1994.

[8] C. Rasmussen and G. Hager. An Adaptive Model for Tracking Objects by Color Alone. Technical Report, DCS-TR-1200, Yale University, 1997.

[9] J. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *ICCV '95*, pp. 612-617, 1995.

[10] D. Reynard, A. Wildenberg, A. Blake, and J. Marchant. Learning Dynamics of Complex Motions from Image Sequences. In *ECCV '96*, pp. 357-368, 1996.

[11] D. Terzopoulos and R. Szeliski. Tracking with Kalman Snakes. In *Active Vision*, A. Blake and A. Yuille, eds., pp. 3-20, 1992.

[12] K. Toyama and G. Hager. Incremental Focus of Attention for Robust Visual Tracking. In *CVPR '96*, pp. 189-195, 1996.

[13] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-Time Tracking of the Human Body. In *SPIE*, Vol. 2615, 1995.