Joint Likelihood Methods for Mitigating Visual Tracking Disturbances

Christopher Rasmussen*
National Institute of Standards and Technology
Gaithersburg, MD
crasmuss@nist.gov

Abstract

We describe a framework that explicitly reasons about data association and combines estimates to improve tracking performance in many difficult visual environments. This work extends two previously reported algorithms: the PDAF, which handles single-target tracking tasks involving agile motions and clutter, and the JPDAF, which shares information between multiple same-modality trackers (such as homogeneous regions, textured regions, or snakes). The capabilities of these methods are improved in two steps: first, by a Joint Likelihood Filter that allows mixed tracker modalities when tracking several objects and accommodates overlaps robustly. A second technique, the Constrained Joint Likelihood Filter, tracks complex objects as conjunctions of cues that are diverse both geometrically (e.g., parts) and qualitatively (e.g., attributes). Rigid and hinge constraints between part trackers and multiple descriptive attributes for individual parts render the whole object more distinctive, reducing susceptibility to mistracking. The generality of our approach allows for easy application to different target types, and it is flexibly defined for straightforward incorporation of other modalities.

1 Introduction

Traditionally, the emphasis in framing visual tracking problems has been on *estimation* [12]. Given a sequence of images containing the object that we wish to represent concisely with a parametric model, an *estimator* is a procedure for finding the parameters of the model which best fit the data. Most of the image data is typically irrelevant, so if the object's image projection can be unambiguously discriminated from the rest of the image, it is segmented and used exclusively for estimation.

Under real world conditions, it can be difficult to accurately identify an object's image projection because visual phenomena such as agile motion, distractions, and occlusions interfere with—"disturb"—estimation. We define *agile motion* as a sustained object movement that exceeds a tracker's dynamic prediction abilities. Its occurrence undermines the estimation process because it renders the putative location of the object's image projection uncertain, complicating efficient segmentation. A further obstacle to clear-cut segmentation is a *distraction*, or another scene element which has a similar image appearance to the object being tracked. Finally, *occlusion* results when another scene element is interposed between the camera and the tracked object, blocking a portion of the object's image projection. All of these factors may bias estimation with bad or missing data; in the worst case, a tracker can lose the target altogether because of them. When tracking multiple similar and/or interacting objects, distractions and occlusions can be particularly problematic.

Clearly, accommodating the visual interactions between single objects and distracting backgrounds as well as among multiple objects requires a combination of both estimation and *correspondence*. By "correspondence" we mean some process for determining what image data to properly associate with an object being tracked and therefore to base the estimation process on. In previous work [14] we adapted to vision two existing *data association* methods: the Probabilistic Data Association Filter (PDAF) and the Joint Probabilistic Data Association Filter (JPDAF) [1]. Our implementations of the PDAF and JPDAF improved tracking performance over standard nearest-neighbor versions of the the Kalman filter for certain classes of visual disturbances: agile single targets with transient distractions and multiple similar (but not overlapping) targets, respectively.

In this paper, we first present a new joint target tracking algorithm, the Joint Likelihood Filter (JLF), which is based on the principles behind the JPDAF but allows for tracked objects to overlap one another and deduces their depth ordering from the image when possible. We also extend the JLF to permit tracked objects to be defined as combinations of geometric parts and qualitative modalities (color, shape, texture, etc.) and therefore more *distinctively*. This method, the Constrained Joint Likelihood Filter (CJLF), tends to mitigate the effect of distractions and occlusions by diversifying the data that estimation is based on and lessening the ambiguity of correspondences.

*This work was carried out while the author was with the Department of Computer Science, Yale University, New Haven, CT

As our algorithms are based on the Kalman filter, they work with point-like *measurements* rather than directly on images. A precursor to both algorithms which we describe here is therefore a sampling process for segmenting and summarizing a discrete set of image areas that resemble the target (where the similarity metric depends on the modality used for tracking). The term "measurement" thus serves as a convenient shorthand for coherent subsets of the image data that may be used for state estimation, and data association serves to weight the influence of these alternatives.

2 Background

At time t, let the state $\mathbf{X}_t \in \mathcal{W}$ represent the current estimate of the tracked object's salient parameters, or state, and let $\mathcal{I}_t = \mathbf{I}_t, \mathbf{I}_{t-1}, \ldots$ be the sequence of images observed so far. Under the Bayesian paradigm, a MAP tracker estimates a state that maximizes $p(\mathbf{X}_t \,|\, \mathcal{I}_t)$. Applying Bayes' theorem and rearranging yields the following expression [1]:

$$p(\mathbf{X}_t \,|\, \mathcal{I}_t) = k_t p(\mathbf{I}_t \,|\, \mathbf{X}_t) p(\mathbf{X}_t \,|\, \mathcal{I}_{t-1}) \quad (1)$$

Here $p(\mathbf{X}_t \mid \mathcal{I}_{t-1})$, which summarizes prior knowledge about $\mathbf{X}_t$, is a prediction based on the previous state estimate and knowledge of the object's dynamics. Asserting that object dynamics are such that states form a Markov chain [5] obtains $p(\mathbf{X}_t \mid \mathcal{I}_{t-1}) = \int_{\mathbf{X}_{t-1}} p(\mathbf{X}_t \mid \mathbf{X}_{t-1}) p(\mathbf{X}_{t-1} \,|\, \mathcal{I}_{t-1})$.

Dropping time indices for brevity, $p(\mathbf{I} \,|\, \mathbf{X})$ describes the probability of observing a particular image at time t given the current state. We call this the *image likelihood*. The image likelihood depends on the physics of image formation and intervening noise [6]. Let π be an *image prediction* function describing the expected image projection of the target given a particular state. If they are not explicitly included in $\mathbf{X}$, assumptions must be made in π about lighting, occlusions, background, object reflectance properties, camera variables such as focal length, etc.

The bases for $p(\mathbf{I} \,|\, \mathbf{X})$ are the form of the predicted target image projection $\pi(\mathbf{X})$ and the method for quantifying the similarity of the image $\mathbf{I}$ to that prediction. Both of these depend on what we call the *modality* used to identify the object. A modality is a visual attribute such as shape, color, direction of motion, etc. that might constitute a tracker's complete description of its target. The three modalities used in this paper—homogeneous (color) regions, textured regions (SSD patches [8]), and snakes [3]—are explicated as modalities in [15]. In non-vision tracking domains such as radar [1], finding peaks in $p(\mathbf{I} \mid \mathbf{X})$ as a measurement-generating precursor to the Kalman filter is fairly simple. A target might be simply a bright point on a dark background, so thresholding alone quickly segments out high-likelihood hypotheses for the target location. Generating visual target measurements, however, is usually more difficult than thresholding and requires more information than just image location. Possible measurement parameters include geometric characteristics such as the location of the area's center and its height, width, and orientation. These parameters define a *measurement space* $\mathcal{Z}$ such that a point $\mathbf{Z} \in \mathcal{Z}$ is related to a state $\mathbf{X}$ via a continuous *measurement function* $H(\mathbf{X}) = \mathbf{Z}$. The measurement function may simply reduce the dimensionality of $\mathbf{X}$ by dropping its temporal parameters, or describe a more complicated relationship between what is measured and what is estimated.

Randomized methods such as the factored sampling approach of the Condensation algorithm [5] have proven successful at finding nonlocal maxima of multimodal image likelihoods. Accordingly, we use a measurement generation method which we call *measurement sampling* which is adapted from factored sampling but retains the notion of locality around a single predicted state. Intuitively, we sample points in state space from the prior distribution on the state $p(\mathbf{X})$, compute their image likelihoods $p(\mathbf{I} \mid \mathbf{X})$, throw away all but the top fraction, and derive the measurement parameters of what remains. Specifically, N samples are taken from a normal distribution with covariance $\mathbf{\Sigma}_{\mathcal{W}}$ in the target's state space $\mathcal{W}$ centered on its current predicted state $\widehat{\mathbf{X}}$. $p(\mathbf{I} \mid \mathbf{X})$ is computed for each sample by scoring the degree of fit between the hypothesized target and the current image. Finally, a winnowing step sorts the samples by their likelihoods and keeps only the n most likely ones ($n \ll N$) for input to the tracking filter. These samples may be "improved" by hill-climbing them using conjugate gradient ascent or Powell's method [12].

3 Joint Likelihood Filter

When there are multiple objects, the form of Eq. 1 becomes more complicated. Assuming conditioning on previous images and letting $\mathcal{X}_T = \mathbf{X}_1, \ldots, \mathbf{X}_T$, we have: $p(\mathcal{X}_T \mid \mathbf{I}) = kp(\mathbf{I} \mid \mathcal{X}_T)p(\mathcal{X}_T)$. The JPDAF works by assorting logically feasible correspondences between targets and measurements. However, it assumes that the first term on the right hand side, which we call the *joint image likelihood*, can be factored as $p(\mathbf{I} \mid \mathcal{X}_T) = p(\mathbf{I} \mid \mathbf{X}_1) \cdots p(\mathbf{I} \mid \mathbf{X}_T)$. Evaluating image likelihoods independently is an approximation that tends to break down when targets abut or overlap because this is exactly when their appearances become dependent on one another. To track objects more accurately, $p(\mathbf{I} \,|\, \mathcal{X}_T)$ must at least consider the ordering of the depths, relative to the camera, of the tracked objects. Knowing which object is in front of which when they overlap is the key to properly predicting the image's appearance $\pi(\mathcal{X}_T)$ from the objects jointly. The JLF attempts to accomplish this by formulating different hypotheses of target orderings and picking the one which makes the current image most likely.

3.1 Joint Measurement Process

The first step in the JLF's *joint measurement* process generates N joint samples. A given joint sample $\mathbf{X}_i^J$, $1 \leq i \leq N$, is built from T *component samples* $\mathbf{X}_j$, $1 \leq j \leq T$, each generated by one of the trackers in its state space $\mathcal{W}_j$ as described in Section 2. The component samples are then stacked to get a joint sample: $\mathbf{X}_i^J = (\mathbf{X}_1, \ldots, \mathbf{X}_T)^T$.

The second step for each joint sample is to pick the most likely depth ordering of its T component samples. To do this, all permutations of depth orderings are enumerated, tagging each component sample with a depth order index in the process. Different depth orderings of non-overlapping component samples are visually equivalent, inducing equivalence classes of depth orderings, so we automatically eliminate all but one representative of each class. Let $\mathbf{D}_{\mathbf{X}_i^J} = \{\mathbf{d}_1, \ldots, \mathbf{d}_{K_{\mathbf{X}_i^J}}\}$ be the set of visually distinct depth order permutations of joint sample $\mathbf{X}_i^J$. For efficiency, we only do gradient ascent on the most likely depth ordering $\mathbf{d}_i$ of each joint sample (a *joint image likelihood* objective function is described in the next subsection) rather than all of them. Finally, the most probable of all of the joint samples $\mathbf{X}_i^J$ is selected and converted to a *joint measurement* $\mathbf{Z}^J$. The component measurements $\mathbf{Z}_1, \ldots, \mathbf{Z}_T$ of $\mathbf{Z}^J$ are then plugged into Kalman filters for their associated trackers.

An example of a joint sample comprising a textured region and a snake is shown in Figure 1(a). The textured region is tracking a chess pawn and the snake is tracking a knight. Since there are two overlapping component samples in the joint sample of the chess example referred to above, there are two depth ordering hypotheses. Hypotheses corresponding to the pawn in front of the knight and the knight in front of the pawn are represented in Figure 1(b) and (e), respectively.

3.2 Joint Image Likelihood

To evaluate the likelihood of a particular joint sample $\mathbf{X}^J$ and its depth ordering $\mathbf{D}_{\mathbf{X}^J}$, the probabilities of its component samples are computed *jointly*. A key difference between this operation and the independent approach of single-object trackers is our ability to predict occlusions between objects. When one object is hypothesized to be in front of another, expectations about the occluded object's appearance change. Trackers of snakes will not expect edges where they are blocked from view, homogeneous region trackers will not expect occluded pixels to fit the color model, and so on. Specifically, $\mathbf{D}_{\mathbf{X}^J}$ allows us to *mask* [7, 16] occluded portions of objects such that the occluding objects take precedence in the formation of a jointly

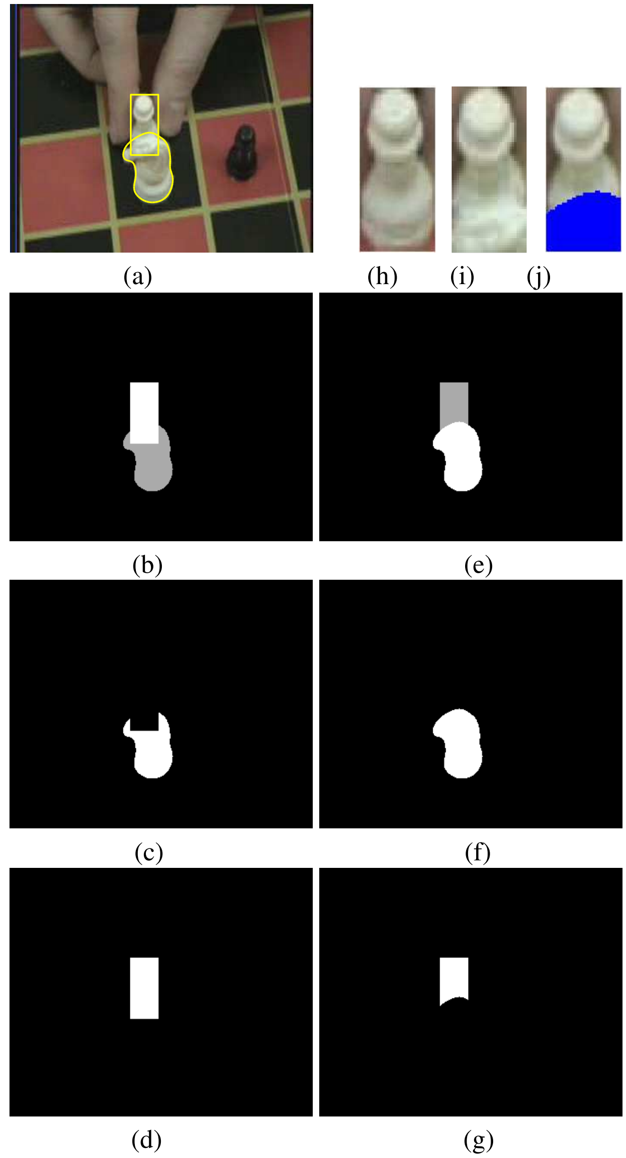


Figure 1. JLF depth orderings. (a) Joint measurement; (b) 1st depth ordering; (c) 1st knight mask; (d) 1st pawn mask; (e) 2nd depth ordering; (f) 2nd knight mask; (g) 2nd pawn mask; (h) Pawn reference image; (i) 1st pawn comparison image; (j) 2nd pawn comparison image.

predicted image $\pi(\mathbf{X}_1, \ldots, \mathbf{X}_T)$. Pixels predicted to be obstructed are ignored and those predicted to be visible are matched normally.

A basic technique of the independent image likelihoods (details of the formulae for homogeneous regions, etc. are given in [15]) is to compute a mean match value ψ over the extent or around the perimeter of the object. Under the JLF, the set of masks $\{\mathbf{M}_j\}$ is used to modify this technique for two reasons. First, some pixels are erroneously counted more than once by single-object trackers when tracked objects overlap; each pixel should only be used as evidence by one tracker. Second, the masks are used to try to ensure that each pixel is counted by the correct tracker. An approach that meets these criteria only counts target pixels that are predicted to be visible in the calculation of that target's mean match value. The masking procedure induced by $\mathbf{D}_{\mathbf{X}^J}$ outputs a binary mask $\mathbf{M}_j$ the size of the image $\mathbf{I}$ for each target t_j. $\mathbf{M}_j(x, y) = 1$ indicates that the image pixel $\mathbf{I}(x, y)$ comes from target t_j and $\mathbf{M}_j(x, y) = 0$ indicates that the pixel belongs to either another object or the background. $\mathbf{M}_{knight}$ is shown for the two depth ordering hypotheses of the chess example in Figure 1(c) and (f). $\mathbf{M}_{pawn}$ is shown for those two hypotheses in Figure 1(d) and (g).

For a textured region t_j, only those interior pixels (x, y) for which $\mathbf{M}_j(x, y) = 1$ contribute to the mean match value. That is, portions of the region's interior that are not visible do not have a match value computed and are subtracted from the effective area. This method is illustrated for the textured-region pawn of the chess example in Figures 1(h),(i), and (j). Figure 1(h) shows the reference image for the pawn. Figures 1(i) and (j) show the comparison images for the hypotheses that the pawn is in front of and behind the knight, respectively. In the latter case, the nearer knight masks out the area of pixels shown in blue. For homogeneous regions, the central area is handled in the same fashion as textured regions, but the inhibitory frame is not in the mask $\mathbf{M}_j$ of the tracker. Rather, only those pixels (x, y) in the inhibitory frame for which $\mathbf{M}_i(x, y) = 0$ for all $i \neq j$ are counted. The same method is also used for snakes: only edges found at locations (x, y) such that $\mathbf{M}_i(x, y) = 0$ for all $i \neq j$ are considered. Finally, any pixels in the interior, frame, or on the normals of an object that are also outside of the image are treated as masked out.

It is also important to guard against interpreting an object as being completely occluded when there is image evidence for its visibility. This problem can be avoided by classifying visible pixels as either positive or negative evidence for the hypothesis that the target is in a certain state, and putting masked pixels in a third, neutral category rather than ignoring them. What makes a pixel a *match*, or positive evidence instead of negative, is fundamentally a threshold Υ in ψ. To quantify this approach, matching pixels are assigned a value of 1, non-matching pixels a value of -1, and masked pixels get 0. Measurements with more corroborative evidence are assigned higher likelihoods than those with no or negative evidence by using the sigmoid function on the sum of the pixel match values.

Specifically, we replace the independent image likelihoods $p(\mathbf{I} \mid \mathbf{X})$ for homogeneous and textured regions and snakes (derived in [15]) with component image likelihoods $p^J(\mathbf{I} \mid \mathbf{X}_j)$. For textured regions, we have:

$$p^J_{tr}(\mathbf{I} \mid \mathbf{X}_j) = \mathrm{sig}\,\Big(\sum_{x,y \in \mathbf{I}_R} a(x, y) \cdot \psi^J_{tr}(x, y)\Big) \qquad (2)$$

$a(x, y)$ is the fraction of the reference patch $\mathbf{I}_R$'s area represented by the pixel at (x, y) and

$$\psi^J_{tr}(x, y) = \begin{cases} 1 & \text{if } \mathbf{M}_j(x, y) = 1 \wedge \\ & (\mathbf{I}_R(x, y) - \mathbf{I}_C(x, y))^2 \leq \Upsilon_{tr} \\ -1 & \text{if } \mathbf{M}_j(x, y) = 1 \wedge \\ & (\mathbf{I}_R(x, y) - \mathbf{I}_C(x, y))^2 > \Upsilon_{tr} \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

where $\mathbf{I}_C$ is the comparison patch in the current image. The component image likelihood's form is analogous for homogeneous regions and snakes; we omit them due to space limitations. More details are in [13].

Let the joint tracker, which has T component trackers, consist of a set $\mathcal{H}$ of homogeneous region trackers, a set $\mathcal{T}$ of textured region trackers, and a set $\mathcal{S}$ of snake trackers such that $T = |\mathcal{H}| + |\mathcal{T}| + |\mathcal{S}|$. The image likelihood of the joint sample $\mathbf{X}^J$ is simply the product of the component likelihoods: $p^J(\mathbf{I} \mid \mathbf{X}^J) = \prod_{t_j \in \mathcal{H}} p^J_{hr}(\mathbf{I} \mid \mathbf{X}_j) \prod_{t_j \in \mathcal{T}} p^J_{tr}(\mathbf{I} \mid \mathbf{X}_j) \prod_{t_j \in \mathcal{S}} p^J_{s}(\mathbf{I} \mid \mathbf{X}_j)$.

4 Constrained Joint Likelihood Filter

We have observed that the more an object is occluded or the better a distracting background feature matches an attribute used for tracking, the more severe the deterioration of accuracy and the greater the chance of outright failure of a PDAF/JPDAF tracker. The approach of the CJLF to the problem of persistent distractors is to try to reduce their incidence, and hence their influence, by defining a target as a conjunction of parts and/or attributes. A tracker with weak discriminatory power can often overcome difficult image conditions because of the *constraints* imposed by its linkage to other trackers. These force consideration of the entire ensemble of parts and attributes simultaneously when interpreting the image, helping to rule out incorrect alternatives.

A linkage between targets means that they are parts of some larger object, and that their states are therefore not independent. This disallows the decomposition of the joint state prior $p(\mathcal{X}_T) = p(\mathbf{X}_1) \cdots p(\mathbf{X}_T)$ that is a vital step in both the JPDAF and JLF. As with the joint image likelihood $p^J(\mathbf{I} \mid \mathbf{X}^J)$ in the previous section, we need a more complex formulation of $p(\mathcal{X}_T)$ that takes into account the interactions between objects by describing how multiple *linked* objects influence one another's states.

The key idea behind the CJLF is an elaboration of one of the most basic kinds of constraints: limitation of the number of parameters in an object's state, which in turn reduces the size of its measurement space. We already use this form of constraint for atomic trackers when we analyze the object, the tracking task, and the visual environment in order to decide what geometric parameters to estimate. To do otherwise only provides the tracker with an opportunity to mistrack along extraneous state dimensions.

The compositional primitives used by the CJLF are based on intuitive physical relationships such as rigid links, hinges, and fixed depth orderings. Given a set of parts or attributes with unconstrained state spaces $\mathcal{W}_1, \ldots, \mathcal{W}_T$, these rules serve as a guide for paring them down to their minimal, constrained forms: $\mathcal{W}'_1, \ldots, \mathcal{W}'_T$. If the paring removes all degrees of freedom of a tracker, its state space becomes empty. Image processing is still performed for that tracker, however, for its contribution to the joint image likelihood.

For purposes of implementation, the CJLF alters the method of obtaining geometric image processing parameters necessary to calculate the image likelihood (such as scale, position, orientation, etc.) whether or not they appear in the state. Let each target t_j have a measurement key $\mathbf{K}_j$ (detailed in [13]). Previously the domain of each function in $\mathbf{K}_j$ was $\mathcal{W}_j$; we now extend it to the joint state space $\mathcal{W}^J$. This allows us to refer to the component measurement geometric parameters of *any* target t_i to define t_j's component measurement geometric parameters. The effect of this reduction in the joint state space is to alter the JLF so that it considers *only* those joint state samples which satisfy the constraints, allowing their joint probabilities to be computed normally. Sampling and hill-climbing can then be used as in the previous section while still meeting the conditions on part relationships.

4.0.1 Constraint Types

Rigid link constraints The simplest kind of constraint between measurements is a *rigid link*. A rigid link between two objects t_1, t_2 implies that t_2's current geometric parameters are completely determined by their initial values and t_1's current values—it has no state or measurement space of its own. Its only function is to contribute to the calculation of the joint image likelihood $p(\mathbf{I} \mid \mathbf{X}_1, \mathbf{X}_2)$. Therefore t_2 does not use a Kalman filter to estimate its own state; its purpose is as an adjunct that makes t_1 a more complex visual object. We denote the rigid link transformation that takes the geometric parameters of object i to those of object j as a function $R_{i,j}$. Thus, $\mathbf{K}_2 = R_{1,2}(\mathbf{K}_1)$ (see [13] for a detailed derivation).

It is straightforward to generalize a two-part, rigidly-constrained joint object to a T-target system. T rigidly-linked parts can be modeled by treating them as $T-1$ linked pairs, every one of which includes target t_1, such that $\mathbf{K}_i = R_{1,i}(\mathbf{K}_1)$.

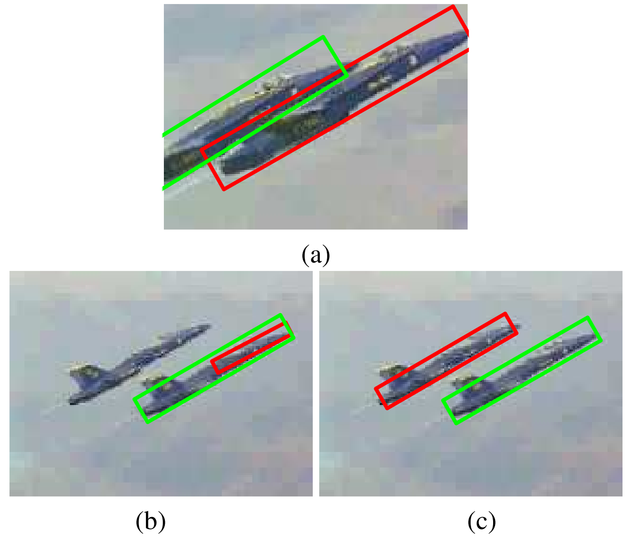


Figure 2. JLF vs. PDAF: Tracking crossing textured regions. (a) Frame 0 for both trackers; (b) PDAF frame 200; (c) JLF frame 200.

Hinge constraints A more complex constraint is a *hinge*, which is like a rigid link but with an angular degree of freedom granted to the second object; the axis of rotation is determined by the initial image location of the hinge: $\bar{x}_h, \bar{y}_h$. The hinge transformation between objects i and j is denoted by $H_{i,j}$.

We can also extend the idea of a single hinge constraint to a *chain* [4] of T parts connected in sequence by $T-1$ hinges. Let C be a chain consisting of T hinge-connected parts: $C = (t_1, \ldots, t_T)$. We can specify the constraint on each part along C inductively: if the first and second links t_1, t_2 are defined by the two-part system introduced above, then the state of the ith part for $i > 1$ is $\mathbf{X}_i = (\phi_i)$ and its measurement space is $\mathcal{Z}_i = \Phi$. Given the measurement key $\mathbf{K}_1$ of the first part t_1, the measurement key of the ith part t_i is given by $\mathbf{K}_i = H_{i-1,i}(H_{i-2,i-1}(\ldots H_{1,2}(\mathbf{K}_1)\ldots))$. By writing $H_{i-1,i}(\mathbf{K}_{i-1})$, the calculations that lead to $\mathbf{K}_{i-1}$ are assumed.

Depth constraints Another useful kind of constraint relates to depth. When there is an expectation that some subset of the objects being tracked will not occlude one another, we can collect them into a *depth group*. Objects in the same depth group are not masked against one another during computation of the joint image likelihood. The primary purpose of depth groups applies to tracking an object with multiple attributes. Since attributes represent qualities of a physical object rather than the object itself, multiple in-

Figure 3. JLF (frames 0, 40, and 80). Deducing the occlusion relationship between a textured region and snake.

stances can be "layered" onto a single object without affecting the visibility of any of them. When a person's face, for example, is tracked by both a textured region tracker (to capture appearance) and a homogeneous region tracker (for skin color), the two trackers are members of the same depth group.

5 Results

Here we give some results for the JLF and CJLF. Input sequences are MPEGs.

5.1 Tracking Objects Jointly

The JLF's superiority over single-object trackers at following crossing objects is illustrated in Figure 2. In this example, two airplanes flying in close formation are tracked using textured regions as they overlap and then separate. The planes scale, translate, and rotate, so the state of each tracker is $\mathbf{X} = (x, y, \phi, s)$, making measurement space $\mathcal{Z} = X \times Y \times \Phi \times S$. For comparison, PDAF trackers assigned to each plane select the best 5 of 250 samples, where the state sampling covariance is $\mathbf{\Sigma}_{\mathcal{W}} = \left(\begin{smallmatrix} 50 & 0 & 0 & 0 \\ 0 & 50 & 0 & 0 \\ 0 & 0 & 0.02 & 0 \\ 0 & 0 & 0 & 0.01 \end{smallmatrix}\right)$. Each of these samples is then improved using Powell's method. The JLF tracker improves the best 5 of 250 joint samples (using $\mathbf{\Sigma}_{\mathcal{W}}$ for each component sample of the joint sample) using Powell's method on the joint image likelihood; the best of these is used to update the state.

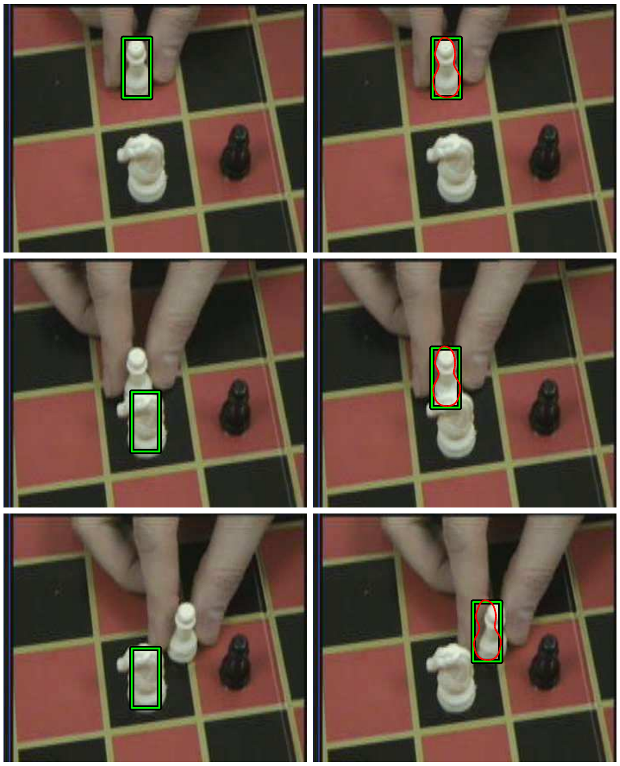
The PDAF plane trackers are attracted by the two nearby matching features in the image as they merge. When the two planes separate, both trackers may follow the same feature, resulting in mistracking. By calculating image likelihoods jointly, the JLF tracker separates properly because of its probabilistic preference for an interpretation that there are two visible objects over an interpretation that one visible object completely occludes the other. The random sampling technique for measurement generation is vital because even using the joint image likelihood, a pure gradient ascent tracker can get stuck in a local minimum as the planes separate. The nonlocality of random sampling allows the trackers to jump out of suboptimal states as the planes separate unambiguously.

The JLF's ability to infer the depth ordering of tracked objects is illustrated in Figure 3. A white pawn chess piece is tracked by a textured region as it moves behind a white knight, which is tracked by a snake; both have state $\mathbf{X} = (x, y, \dot{x}, \dot{y})$, making each component's measurement space $\mathcal{Z} = X \times Y$. Measurement generation uses pure gradient ascent with Powell's method. A tracker's outline, normally colored, is drawn in gray when the most likely depth ordering indicates that it is partially occluded. The fact that the pawn is behind the knight during the middle section of the tracking sequence is correctly deduced.

5.2 Tracking Objects with Constraints

Now suppose we want to track the pawn from the previous example without knowing about the knight. Using a JLF tracker consisting solely of a homogeneous region initialized as shown in frame 0 of Figure 4(a), the state is $\mathbf{X} = (x, y)$, making measurement space $\mathcal{Z} = X \times Y$. The single most likely of 50 samples from a sampling covariance of $\mathbf{\Sigma}_{\mathcal{W}} = \left(\begin{smallmatrix} 50 & 0 \\ 0 & 50 \end{smallmatrix}\right)$ is improved with Powell's method. This approach fails because the untracked white knight fits the color model well and attracts the pawn strongly. The fundamental problem is the presence of a strong, persistent peak due to the knight in the homogeneous region's image likelihood that is not expected by the JLF tracker.

Tracking the pawn in a similar fashion with a snake alone yields better results because $p_s(\mathbf{I} \mid \mathbf{X})$ has only one prominent extremum rather than two. This quantifies our intuition that shape is a better cue for this task than color. Without knowing ahead of time which modality, if any, is sufficiently distinctive for successful tracking, a prudent strategy is to use multiple attributes simultaneously. The conjunction of



(a) (b)

Figure 4. Multi-attribute CJLF (frames 0, 50, and 100). (a) JLF homogeneous region tracker is distracted by the white knight; (b) CJLF homogeneous region and snake tracker overcomes the distraction.

color and shape in one depth group results in a joint image likelihood $p^J(\mathbf{I} \mid \mathbf{X}^J)$ with peaks only where *both* likelihoods $p_{hr}(\mathbf{I} \mid \mathbf{X})$ and $p_s(\mathbf{I} \mid \mathbf{X})$ have peaks, reducing distractions. Formally, we utilize the pawn's color and shape simultaneously by modeling it with two rigidly-linked attributes: a homogeneous region and a snake with coincident centers. The pawn's joint region-snake tracker follows the same regime of hill-climbing on the single best of 50 samples as the single-attribute trackers above. As Figure 4(b) shows, this constrained formulation permits the pawn to be successfully tracked when the homogeneous region alone fails.

A more complicated situation which shows the advantage of the CJLF over the JLF is shown in Figure 5. Here we want to track a person's hand and forearm as homogeneous regions while they shake hands with another person, who is not tracked. Each component ($i = 1, 2$) of the JLF tracker has a state of the form $\mathbf{X}_i = (x_i, y_i, \phi_i, \dot{x}_i, \dot{y}_i, \dot{\phi}_i)$ with measurement spaces $\mathcal{Z}_1 = \mathcal{Z}_2 = X \times Y \times \Phi$. Accelerations during the handshake are too large for pure gradient tracking, so each component tracker selects the best 1 of 50 samples, where $\mathbf{\Sigma}_{\mathcal{W}} = \left(\begin{smallmatrix} 50 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 0 & 0.02 \end{smallmatrix}\right)$ and hill-climbs it using Powell's method. Despite these measures, the hand tracker mistracks when its target is in close proximity to the other person's hands, and the forearm tracker erroneously slides along the sleeve.

These shortcomings can be eliminated with a hinge constraint joining the hand and forearm trackers at the midpoints of their abutting short sides. The state of the forearm tracker remains the same, while the hand tracker is reduced to one degree of angular freedom. Adopting this approach prevents the hand and forearm trackers from floating apart; relatively higher joint image likelihoods keep the hinge at the sleeve-hand border. The result is that during the period of ambiguity when the two hands are clasped, a realistic interpretation of the situation is maintained and tracking proceeds correctly after the hands separate.

6 Related Work

Most of the previous work on tracking complex objects has not explicitly tackled the data association issue. One line of primarily motion-based tracking work has avoided the problem through a differential approach. For example, Yamamoto et al. [17], tracked in-plane articulated movements of a human arm by relating arm motion to image change via the Jacobian and solving the brightness equation using least-squares. Basu et al. [2] used a similar technique to recover 3-D head motion parameters, and other assemblages of body parts have been tracked in [16, 11, 10, 4]. These efforts are closer to pure estimation, rather than the simultaneous problem of estimation and explicit label assignment that we focus on.

The Condensation algorithm [5] combats clutter by maintaining a set of hypotheses about associations and resolving them over time. It uses random sampling as well, but does not have an explicit notion of state. Rather, the samples must be queried to obtain one. The suggested query procedure does not work well when the image likelihood is multimodal, and the authors note that a more sophisticated "mode finder" is necessary. This is essentially what our measurement generation algorithm of Section 3.1 implements.

The difficulties arising from mutual occlusions among tracked objects have been addressed for flexing human fingers in [16] and cars on a highway in [7]. The essential idea of both approaches was to mask out the occluded part to prevent it from claiming the measurement generated by the occluding part. In each case, 3-D information was available in order to predict which part was occluded, though recently an image-based, occlusion-handling extension to the Condensation algorithm for snakes was described in [9].

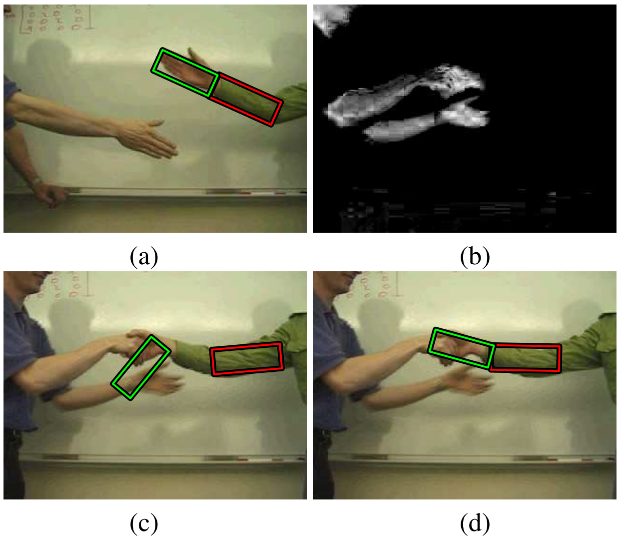


Figure 5. CJLF with geometric constraints. (a) Frame 0: homogeneous region trackers on hand and forearm; (b) Frame 260: hand color similarity; (c) Frame 260: single-object hand tracker is distracted by other person's hand; (d) Frame 260: hinge constraint imposed by CJLF permits accurate tracking of the hand.

7 Conclusion

This paper's primary contribution is its demonstration of the importance of reasoning about correspondences between trackers and image data in order to achieve robust vision-based tracking in the presence of visual disturbances. We explicated shortcomings in the JPDAF and remedied them with a more efficient and sophisticated method, the JLF. Our extension of this method to collections of objects of different modalities such as color, shape, and appearance is original. Moreover, though others have used three-dimensional state parameters to assist with occlusion reasoning, the JLF's inference of the depth ordering of tracked objects from image data is novel.

Finally, we augmented the JLF method to allow low-level trackers to be composed via part and attribute constraints in order to specify more complex targets. This algorithm, the CJLF, reduces the vulnerability of a vision-based tracker to unmodeled distractions and occlusions by effectively defining its target more distinctively. Although geometric constraints are a well-established method for increasing robustness, exploiting multiple modalities simultaneously to track a single object—especially three, as we do—is fairly new, and the union of these two approaches is clearly an advance.

References

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

[2] S. Basu, I. Essa, and A. Pentland. Motion regularization for model-based head tracking. In *Proc. Int. Conf. Pattern Recognition*, 1996.

[3] A. Blake, M. Isard, and D. Reynard. Learning to track the visual motion of contours. *Artificial Intelligence*, (78):101–133, 1995.

[4] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proc. Computer Vision and Pattern Recognition*, pages 8–15, 1998.

[5] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *Proc. European Conf. Computer Vision*, pages 343–356, 1996.

[6] D. Knill, D. Kersten, and A. Yuille. Introduction: A Bayesian formulation of visual perception. In D. Knill and W. Richards, editors, *Perception as Bayesian Inference*, pages 1–21. Cambridge University Press, 1996.

[7] D. Koller, J. Weber, and J. Malik. Robust multiple car tracking with occlusion reasoning. In *Proc. Computer Vision and Pattern Recognition*, pages 189–196, 1994.

[8] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[9] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. In *Proc. Int. Conf. Computer Vision*, 1999.

[10] D. Morris and J. Rehg. Singularity analysis for articulated object tracking. In *Proc. Computer Vision and Pattern Recognition*, pages 289–296, 1998.

[11] A. Pentland and B. Horowitz. Recovery of nonrigid motion and structure. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(7):730–742, July 1991.

[12] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1993.

[13] C. Rasmussen. *Integrating Multiple Visual Cues for Robust Tracking*. PhD thesis, Yale University, New Haven, CT, 2000.

[14] C. Rasmussen and G. Hager. Joint probabilistic techniques for tracking multi-part objects. In *Proc. Computer Vision and Pattern Recognition*, pages 16–21, 1998.

[15] C. Rasmussen and G. Hager. Probabilistic data association methods for tracking multiple and compound visual objects. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2001. Under review.

[16] J. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *Proc. Int. Conf. Computer Vision*, pages 612–617, 1995.

[17] M. Yamamoto and K. Koshikawa. Human motion analysis based on a robot arm model. In *Proc. Computer Vision and Pattern Recognition*, 1991.