Multimodal Partial Estimates Fusion

Jiang Xu, Junsong Yuan, Ying Wu
Department of Electrical Engineering and Computer Science, Northwestern University
2145 Sheridan Road, Evanston, IL 60208, USA
jxu323@eecs.northwestern.edu, j-yuan@u.northwestern.edu, yingwu@eecs.northwestern.edu

Abstract

Fusing partial estimates is a critical and common problem in many computer vision tasks such as part-based detection and tracking. It generally becomes complicated and intractable when there are a large number of multimodal partial estimates, and thus it is desirable to find an effective and scalable fusion method to integrate these partial estimates. This paper presents a novel and effective approach to fusing multimodal partial estimates in a principled way. In this new approach, fusion is related to a computational geometry problem of finding the minimum-volume orthotope, and an effective and scalable branch and bound search algorithm is designed to obtain the global optimal solution. Experiments on tracking articulated objects and occluded objects show the effectiveness of the proposed approach.

1. Introduction

Many computer vision tasks involve the estimation of the unknown $\mathbf{x} \in \mathbb{R}^d$ from many independent estimates $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$, where the individual estimate $\mathbf{y}_i$ may be obtained from various sources (*e.g.*, different views, time and cues), or from partial features. We refer each individual estimate as the *partial estimate* (PE) or *partial belief*, and the final estimation as the *complete estimation* (CE). A PE $\mathbf{y}_i$ gives an individual estimate of the unknown $\mathbf{x}$, and it may only provide the estimate on several specific dimensions of $\mathbf{x}$, so it is called a partial estimate. As the PEs can be quite inaccurate, a critical question is how we can fuse these partial estimates for a better estimation, *i.e.*, how to obtain $\mathbf{x} = \text{fuse}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$.

One concrete example is the part-based object detection and tracking. The target is represented by its parts and each part is associated with a dedicated detector and tracker, each of which provides a PE of the location and motion of the target. Because a part of the target is generally less discriminative than the entire target, the matching to this part is likely to include many false positives. This is especially true when the target is in a clutter background. Therefore, its PEs tend to have multiple modes, where most of them correspond to false positive matches. We refer one such PE that has multiple modes as a *multimodal partial estimate* (MPE), and our work is focused on the fusion of the MPEs. We want to emphasis that in our work we refer the word *multimodal* to multiple modes in *one* PE, rather than multiple estimates/sensors (Fig. 1).

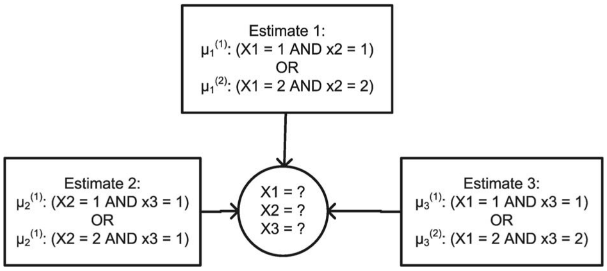
Estimate 1:
$\mu_1^{(1)}$: (X1 = 1 AND x2 = 1)
OR
$\mu_1^{(2)}$: (X1 = 2 AND x2 = 2)
Estimate 2:
$\mu_2^{(1)}$: (X2 = 1 AND x3 = 1)
OR
$\mu_2^{(1)}$: (X2 = 2 AND x3 = 1)
X1 = ?
X2 = ?
X3 = ?
Estimate 3:
$\mu_3^{(1)}$: (X1 = 1 AND x3 = 1)
OR
$\mu_3^{(2)}$: (X1 = 2 AND x3 = 2)

Figure 1. Example of fusing three MPEs, and each MPE has two modes. The optimal fusion result CE is the most consistent **x** with all the three MPEs.

If the PEs $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$ are all unimodal, it is possible to obtain a closed-form fusion for the CE $\mathbf{x}$, *e.g.*, through the best linear unbiased estimation (BLUE) [10]. However, when $\mathbf{y}_i$ is multimodal (*e.g.*, modeled as a Gaussian mixture), the fusion for the CE is likely to exhibit an extremely complicated form. If each MPE has m modes, the number of modes in the CE is in the order of $o(m^n)$. In its discrete case, suppose each MPE consists of a set of m discrete estimates, the complexity of searching for the best CE shall be $o(m^n)$. Such an exponential growth of the number of modes (or the combinatorial complexity in the discrete case) makes any form of fuse$(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$ very difficult to be optimized. As the complicated CE has an enormous number of local optima, fusion is likely to end up with a low-quality estimation unless an exhaustive search can be performed. Even when an exhaustive search is merely viable when n is small, such a method is not scalable when there are many MPEs to fuse. Thus, new scalable fusion methods are desirable.

As the CE $\mathbf{x} = \text{fuse}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$ in general may not have good analytical properties, it is difficult to manipulate it directly. In this paper, we convert the error minimization in the fusion problem into a problem that finds a minimum-volume d-orthotope in $\mathbb{R}^d$ subject to some constraints. The *minimum-volume orthotope* problem can also be viewed as a multi-class generalization of the closest-pair problem in computational geometry. We design an effective branch and bound search algorithm to determine the global optimal solution to this problem with a moderate computational complexity.

The novelty of this work includes the following three aspects. (1) The fusion of MPEs is converted to a tractable minimum-volume orthotope problem, in which the intricate CE is exactly optimized in a discrete view, or approximately optimized in a continuous view. This new treatment leads to a tractable solution to fusion. (2) It reveals an interesting connection between probabilistic data fusion and computational geometry. The proposed solution to the minimum-volume orthotope problem provides a non-trivial generalization of the closest-pair problem. (3) The proposed fusion method is very scalable w.r.t. the number of estimates, or information sources, as the complexity is almost constant w.r.t. the number of sources.

2. Related Work

There have been extensive studies on distributed estimates fusion. In [10], two criteria for optimal fusion of unimodal Gaussian estimates are summarized. One is the weighted least squares (WLS), and the other is the best linear unbiased estimation (BLUE). These fusion techniques can be applied to some classic computer vision problems such as the optical flow estimation [11].

When bad or fault estimates exist, WLS or BLUE cannot work well. To better handle noisy estimates, which refer to bad or outlier estimates, one possible solution is to allocate large variances to the bad estimates, such as in the methods of Covariance Intersection/Union (CI/CU) [13]. In [17], the fusion problem when the measurement errors are heteroscedastic is addressed, and the problem is solved in a WLS way. Another solution to handle noisy estimates is to keep good estimates while discarding bad ones. Variable-Bandwidth Density-based Fusion (VBDF) [2] falls into such a category, which performs globally, and attempts to alleviate the influence of the outliers by gradually reducing the bandwidth of the modes. By applying VBDF, a tracking method is presented in [4]. However, it cannot guarantee the global optimality in fusion. Another method to alleviate the effect of bad estimates is to measure the goodness of the estimates locally. For example in [6], a principle to estimate the fidelity of each measurement in a localized calculation is presented.

Unfortunately, all of the above fusion methods are not designed for multimodal cases, namely multiple modes in one estimate. To handle the multimodal estimates fusion problem, there have been two types of solutions: using distributed algorithms [15, 12], or using randomized algorithms [3]. If the estimates can be represented in a loosely-connected graph, several techniques can be applied, such as the variational methods [7], Belief Propagation (BP) [15], or Nonparametric Belief Propagation (NBP) [12]. However, if the graph is densely connected, these methods are easily trapped by local minima, or cannot even converge, due to the loops in the densely connected graph. To avoid local minima and to guarantee the convergence, randomized algorithms can be applied. For example, RANSAC [3] has the ability to obtain a robust estimation from noisy MPEs. Even if only one mode is correct, and all of the others are outliers in each MPE, RANSAC may still obtain the global optimum with some probability, but the performance of RANSAC deteriorates when the number of modes in each MPE or the number of the MPEs increases.

3. Problem Formulation and Solution

Given a collection of MPEs $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$, we want to obtain the CE of the high dimensional unknown $\mathbf{x} \in \mathbb{R}^d$. To better explain our idea, we first examine the problem of fusing discrete MPEs, then present the solution of fusing continuous MPEs in Sec. 4.2.

For discrete MPEs, each MPE $\mathbf{y}_i$ contains multiple point estimations (*i.e.* modes):

$$\mathbf{y}_i = \{\mathbf{y}_i^1, \cdots, \mathbf{y}_i^{\nu_i}\},$$

where $\mathbf{y}_i^j \in \mathbb{R}^{h_i}$ is the j-th mode of $\mathbf{y}_i$ in the h_i-dimensional subspace ($h_i \leq d$), and ν_i is the number of modes in $\mathbf{y}_i$. Given two modes belonging to two different subspaces, $\alpha \in \mathbb{R}^A$ and $\beta \in \mathbb{R}^B$, we define their arithmetic operations (summation, subtraction and maximization) as below.

Computation Rule 1. Addition and Subtraction

In addition or subtraction, we only perform the calculation in space $\mathbb{R}^A \bigcap \mathbb{R}^B$, i.e.*, if $\gamma = \alpha \pm \beta$, then*

$$\gamma_i = \begin{cases} \alpha_i \pm \beta_i & \textit{if } i \in \mathbb{R}^A \bigcap \mathbb{R}^B \\ \textit{undefined} & \textit{otherwise} \end{cases}$$

Computation Rule 2. Max and Min

In maximization or minimization, we perform the calculation in space $\mathbb{R}^A \bigcup \mathbb{R}^B$, i.e.*, if $\gamma = \max(\alpha, \beta)$, then*

$$\gamma_i = \begin{cases} \max(\alpha_i, \beta_i) & \textit{if } i \in \mathbb{R}^A \bigcap \mathbb{R}^B \\ \alpha_i & \textit{if } i \in \mathbb{R}^A \bigcap \mathbb{R}^{\overline{B}} \\ \beta_i & \textit{if } i \in \mathbb{R}^{\overline{A}} \bigcap \mathbb{R}^B \\ \textit{undefined} & \textit{otherwise} \end{cases}$$

An illustrative example is shown in Fig. 2.

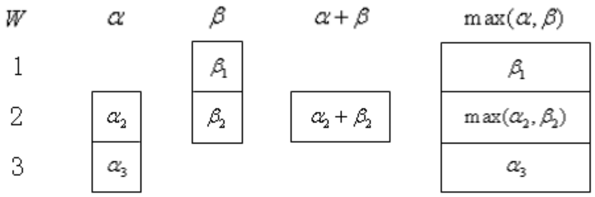

Figure 2. An illustration of arithmetic operations over two vectors belonging to two different subspaces. Suppose there are two vectors $\alpha \in \mathbb{R}^A$ and $\beta \in \mathbb{R}^B$. We use α_i to denote the value at α's i-th dimension. For example, here $A = \{2, 3\}$, $B = \{1, 2\}$, which means α is in the 2nd and 3rd dimensions of the whole space, and β is in the 1st and 2nd dimensions of the whole space. As the figure shows, $\alpha = [11, 12]$, $\beta = [13, 14]$, then $\alpha_2 = 11$, $\alpha_3 = 12$, $\beta_1 = 13$, $\beta_2 = 14$, while α_1 and β_3 are undefined. The addition is performed at the intersection of the subspaces, while the maximization is performed at the union of the subspaces.

3.1. The Objective Function

A good fusion result should be consistent with the MPEs. One natural objective is to minimize the average estimation error, *i.e.*,

$$\min_{\mathbf{x}} \frac{1}{n} \sum_i \Psi(\mathbf{x}, \mathbf{y}_i), \tag{1}$$

where $\Psi(\mathbf{x}, \mathbf{y}_i)$ is the measurement of the inconsistency. For example, we can choose $\Psi(\mathbf{x}, \mathbf{y}_i)$ as:

$$\Psi(\mathbf{x}, \mathbf{y}_i) = \min_j \|\mathbf{x} - \mathbf{y}_i^j\|_\infty,$$

where the final estimation $\mathbf{x}$ is expected to be consistent with at least one of the modes of $\mathbf{y}_i$. Here the L_∞ norm of a vector α is $\|\alpha\|_\infty = \max_i |\alpha_i|$. Although other types of measurements are possible, we will show later that the L_∞ norm leads to an elegant global optimal solution.

Considering that it is difficult to minimize the average estimation error in Eq. 1, we slightly change the original formulation by replacing the average error with the maximum error among the MPEs:

$$\mathbf{x}^* = \arg_{\mathbf{x}} \min_{\mathbf{x}} \max_i \Psi(\mathbf{x}, \mathbf{y}_i), \tag{2}$$

or the median error among the MPEs:

$$\mathbf{x}^* = \arg_{\mathbf{x}} \min_{\mathbf{x}} median_i \Psi(\mathbf{x}, \mathbf{y}_i). \tag{3}$$

We call Eq.2 as the *maximum fusion* while Eq. 3 as the *median fusion*. The median fusion is less sensitive to noise while the maximum fusion may be influenced by an outlier MPE, in which all its modes are outliers. For clarity, we mainly discuss how to solve the maximum fusion in this section. The solution to the median fusion follows the same strategy and will be briefly discussed in Sec. 3.4 as well.

3.2. Equivalence to Orthotope Search

The minimization of Eq. 2 can be converted into a min-volume orthotope search problem, as explained in Figure 3.

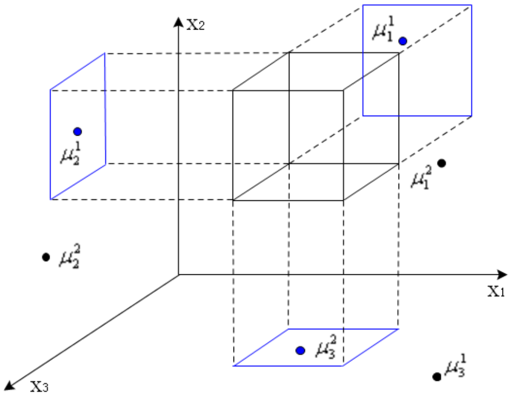

Figure 3. An illustration of orthotope search for partial estimation fusion. There are three MPEs, $\mathbf{y}_1 = \{\mu_1^1, \mu_1^2\}$, $\mathbf{y}_2 = \{\mu_2^1, \mu_2^2\}$, $\mathbf{y}_3 = \{\mu_3^1, \mu_3^2\}$. The orthotope V contains a mode if and only if this mode is contained in V's projection to this mode's subspace, e.g, $\mu_1^1 \in V$ and $\mu_3^1 \notin V$. Minimizing estimation error (Eq. 2) is equivalent to minimizing the volume of the orthotope V, which must contain at least one mode from each MPE (Eq. 4). In this figure, the orthotope V contains μ_1^1, μ_2^1, μ_3^2, so $\mathcal{W}(V) = 1$.

Our task is to find an orthotope (a high-dimensional bounding box) that can cover at least one mode from every MPE $\mathbf{y}_i$. To minimize the maximum error in Eq. 2, we require the longest edge of the orthotope has the minimum length. Based on the above definitions, we propose the following optimization problem

$$\begin{aligned} \min \quad & \|V\|_\infty \\ \text{s.t} \quad & \mathcal{W}(V) = 1. \end{aligned} \tag{4}$$

Here we denote by V a d-dimensional axis-aligned orthotope. An orthotope V's **volume** is related to the length of its longest edge, denoted by $\|V\|_\infty$. $\mathcal{W}(V)$ is the predicate function of the orthotope V:

$$\mathcal{W}(V) = \begin{cases} 1 & \forall\, i, \exists\, j, \text{such that } \mathbf{y}_i^j \in V \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

When a lower-dimensional mode $\mathbf{y}_i^j$ is inside the subspace projection of the d-dimensional axis-aligned orthotope V, the orthotope V **contains** the mode $\mathbf{y}_i^j$, and we denote by $\mathbf{y}_i^j \in V$ (Figure 3).

To justify our formulation as an orthotope search problem, we prove the equivalence between Eq. 2 and Eq. 4 in Theorem 1. We further derive the property under the condition of unique optimal solution in Theorem 2. The proof of both theorems can be found in [14].

Theorem 1. *The equivalence of the optimization in Eq. 2 and Eq. 4*

Let

$$v_1 = \min_{\mathbf{x}} \max_i \Psi(\mathbf{x}, \mathbf{y}_i)$$

and

$$v_2 = \min_V \|V\|_\infty, \qquad \textit{s.t.} \quad \mathcal{W}(V) = 1.$$

Then

$$v_1 = v_2/2.$$

Theorem 2. *If Eq. 2 has a unique optimal $\mathbf{x}^*$ and V^* is the optimal solution to Eq. 4, then $\mathbf{x}^*$ is the center of V^*.*

3.3. A Branch and Bound Solution

According to Theorem 2, we solve Eq. 2 by optimizing Eq. 4, which is to find a minimum-volume orthotope satisfying the predicate. In order to obtain the global optimal solution in the high-dimensional space, we propose a branch and bound search algorithm to find the best orthotope efficiently. As an efficient search method, branch and bound has been applied to object detection [9] and action detection [16]. Our solution is related to [9, 16], but works in a high-dimensional discretized space.

Algorithm 1: Maximum Fusion of MPEs

```
input  : Multimodal partial estimates (MPEs)
         $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$
output: Complete estimation (CE) $\mathbf{x}$

Initialize $\mathbb{V}$ as the collection of all orthotope
candidates in the $d$-dimensional space.
Initialize an empty priority queue $\mathcal{Q}$, in which the
element with the smallest key value pops first.

repeat
    split $\mathbb{V}$ into $\mathbb{V}'$ and $\mathbb{V}''$
    if $\mathcal{W}(\overline{\mathbb{V}'}) = 1$ then
        $\mathbb{V}' \rightarrow \mathcal{Q}$ by the key value $\|\underline{\mathbb{V}'}\|_\infty$
    if $\mathcal{W}(\overline{\mathbb{V}''}) = 1$ then
        $\mathbb{V}'' \rightarrow \mathcal{Q}$ by the key value $\|\underline{\mathbb{V}''}\|_\infty$
    retrieve the top element $\mathbb{V}$ from $\mathcal{Q}$
until $\mathbb{V}$ contains only one element
retrieve the only element $V^*$ of $\mathbb{V}$
return $\mathbf{x}$ as the center point of $V^*$
```

Our branch and bound search algorithm is presented in Algorithm 1. Let $\mathbb{V} = \{V_i\}$ be an *orthotope-set*, where each V_i is an orthotope in the d-dimensional space. The union of $\mathbb{V}$, denoted by $\overline{\mathbb{V}}$, is the minimum orthotope which satisfies $\forall V \in \mathbb{V}, \overline{\mathbb{V}} \supseteq V$. The intersection of $\mathbb{V}$, denoted by $\underline{\mathbb{V}}$, is the maximum orthotope which satisfies $\forall V \in \mathbb{V}, \underline{\mathbb{V}} \subseteq V$. We provide an illustrative example in Figure 4.

Given the original orthotope-set $\mathbb{V}$, our task is to find a minimum-volume $V^* \in \mathbb{V}$ satisfying the predicate, and the optimal CE $\mathbf{x} \in \mathbb{R}^d$ can be uniquely determined by V^*. In each iteration in Algorithm 1, we split $\mathbb{V}$ into two parts $\mathbb{V}'$ and $\mathbb{V}''$, and the splitting point is the middle point of the longest dimension of $\mathbb{V}$ in the orthotope-set space.

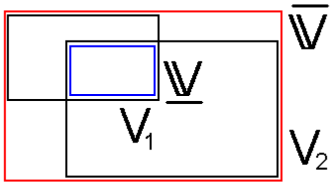

Figure 4. An example showing the upper and lower bounds of $\mathbb{V}$. The two black rectangles are V_1 and V_2, and $\mathbb{V}$ = $\{V_1, V_2\}$. $\overline{\mathbb{V}}$ is the red rectangle which contains V_1 and V_2, and $\underline{\mathbb{V}}$ is the blue rectangle which is contained by V_1 and V_2.

We use the bound properties of $\overline{\mathbb{V}}$ and $\underline{\mathbb{V}}$: If the union of $\mathbb{V}$ cannot satisfy the predicate, it impossible for any $V \in \mathbb{V}$ to satisfy the predicate. As a result, only if $\overline{\mathbb{V}}$ satisfies the predicate, it is worth to perform a further check on $\mathbb{V}$. Otherwise this $\mathbb{V}$ can be safely pruned.

We index $\mathbb{V}$ with a key value $\|\underline{\mathbb{V}}\|_\infty$. This key value provides a lower bound, *i.e.* $\forall V \in \mathbb{V}, \|\underline{\mathbb{V}}\|_\infty \leq \|V\|_\infty$. We use a priority queue $\mathcal{Q}$ to store the orthotope-sets by their key values. Each time we retrieve from $\mathcal{Q}$ a candidate orthotope-set $\mathbb{V}$ with the smallest key value. The retrieving process keeps going until the retrieved $\mathbb{V}$ contains only one orthotope V^*, then V^* has to be the optimal solution because V^* satisfies the predicate and has a minimum volume compared with all other possible orthotopes in $\mathcal{Q}$.

3.4. Median Fusion

As mentioned earlier, the limitation of the maximum fusion is that it is sensitive to noisy MPEs. To address this issue, we can modify our objective function to a robust form by using the median fusion in Eq.3. The corresponding orthotope search problem can still be formulated in Eq. 4, but with a different predicate function:

$$\mathcal{W}(V) = \begin{cases} 1 & \text{for at least half of } i, \exists\, j, \text{such that } \mathbf{y}_i^j \in V \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Using $\mathcal{W}(V)$ in Eq. 6 and the same branch and bound procedure as in Algorithm 1, we can obtain the optimal solution to the median fusion.

4. Beyond Basic Formulation

Although we obtain the global optimal solution under the discrete MPE case, the MPE fusion is more difficult when each MPE provides a continuous estimation. In this section, we firstly show the connection between our algorithm and the bichromatic pair problem in computational geometry, then we extend our solution to the continuous MPE fusion and provide a probabilistic interpretation of our approach.

4.1. Link to Computational Geometry

The bichromatic pair problem [8] is formulated as

$$\min_{j(1),j(2)} |\mathbf{y}_1^{j(1)} - \mathbf{y}_2^{j(2)}|, \tag{7}$$

where the objective is to find the closest pair $\mathbf{y}_1^{j(1)} \in \mathbf{y}_1$ and $\mathbf{y}_2^{j(2)} \in \mathbf{y}_2$ from different classes $\mathbf{y}_1$ and $\mathbf{y}_2$.

We extend this problem to multiple classes, as well as multiple subspaces, where $\mathbf{y}_i$ and $\mathbf{y}_k$ are two MPEs and can belong to different subspaces. The *multichromatic pair problem* is similar to Eq. 7:

$$\min_{j(\cdot)} \max_{\text{for all } i,\, k} \|\mathbf{y}_i^{j(i)} - \mathbf{y}_k^{j(k)}\|_\infty, \tag{8}$$

where the goal is to find a mode from each MPE, such that the maximum distance among all mode-pairs is minimized. Accroding to the following Theorem, the multichromatic pair problem is equivalent to Eq. 4, therefore it can be solved by our proposed branch and bound method as well.

Theorem 3. *The equivalence of Eq. 4 and Eq. 8.*
Let

$$v_2 = \min \|V\|_\infty, \qquad s.t. \quad \mathcal{W}(V) = 1.$$

and

$$v_3 = \min_{j(\cdot)} \max_{\text{for all } i,\, k} \|\mathbf{y}_i^{j(i)} - \mathbf{y}_k^{j(k)}\|_\infty.$$

Then

$$v_2 = v_3.$$

The proof of Theorem 3 can be found in [14].

Corollary 1. *Optimizing Eq. 2, Eq. 4 and Eq. 8 are equivalent.*

In summary, fusing discrete MPEs can be converted to finding a minimum orthotope containing at least one mode from each MPE, and is also equivalent to the multichromatic pair problem.

4.2. MPE Fusion in a Probabilistic View

Now we consider the fusion of continuous MPEs. Suppose that each MPE $\mathbf{y}_i$ generates a multimodal distribution $p_i(\mathbf{x}|\mathbf{y}_i)$:

$$p_i(\mathbf{x}|\mathbf{y}_i) = \sum_j p(\mathbf{y}_i^j) k(\mathbf{x} - \mathbf{y}_i^j), \tag{9}$$

where $p(\mathbf{y}_i^j)$ is the prior of mode $\mathbf{y}_i^j$. If $k_i(\cdot)$ is the Gaussian kernel, then $p_i(\mathbf{x}|\mathbf{y}_i)$ is a Gaussian Mixture (GM). In our definition of $k_i(\cdot)$, we call $p_i(\mathbf{x}|\mathbf{y}_i)$ an Infinity Mixture (IM), as $k_i(\cdot)$ uses the L_∞ norm:

$$k_i(\alpha) = C_i \exp(-\frac{\|\alpha\|_\infty}{\sigma}), \tag{10}$$

where σ is the kernel bandwidth, and C_i is the normalization term. This IM justifies our previous optimization method in a probabilistic view.

Denote by $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$. Suppose $p(\mathbf{x}|\mathbf{Y})$ follows the Products of Experts (PoE) model [5], the distribution becomes:

$$p(\mathbf{x}|\mathbf{Y}) \propto \prod_i p_i(\mathbf{x}|\mathbf{y}_i), \tag{11}$$

where $p_i(\mathbf{x}|\mathbf{y}_i)$ is the partial estimation, or partial belief of $\mathbf{x}$ from $\mathbf{y}_i$, and we assume that $p_i(\mathbf{x}|\mathbf{y}_i)$ are independent. Our objective is to find an estimate $\mathbf{x} \in \mathbb{R}^d$ with the highest probability:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{Y}), \qquad \mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n\}. \tag{12}$$

Searching $p(\mathbf{x}|\mathbf{Y})$ in the high dimensional space is an extremely difficult problem. For an arbitrary $\mathbf{x} \in \mathbb{R}^d$, we consider the orthotope V **centered at** $\mathbf{x}$. If we only count the modes located inside the orthotope, while ignoring the modes outside the orthotope, we obtain the following lower bound of $p(\mathbf{x}|\mathbf{Y})$ by combining Eq. 9, 10 and 11:

$$\begin{aligned} p(\mathbf{x}|\mathbf{Y}) &= \mathcal{C}_1 \prod_i \sum_j p(\mathbf{y}_i^j) k(\mathbf{x} - \mathbf{y}_i^j) \\ &\geq \mathcal{C}_1 \prod_i \sum_{\hat{j}} p(\mathbf{y}_i^j) k(\mathbf{x} - \mathbf{y}_i^j) \\ &= \mathcal{C}_1 \prod_i \sum_{\hat{j}} p(\mathbf{y}_i^j) C_i \exp(-\frac{\|\mathbf{x} - \mathbf{y}_i^j\|_\infty}{\sigma}) \\ &\geq \mathcal{C}_1 \prod_i \sum_{\hat{j}} p(\mathbf{y}_i^j) C_i \exp(-\frac{\|V\|_\infty}{2\sigma}) \end{aligned}$$

where $\hat{j}$ only counts the modes inside V, and $\mathcal{C}_1$ is a constant. The first inequality is obtained by ignoring the contribution from the modes outside the orthotope, and the second inequality is obtained from $\|\mathbf{x} - \mathbf{y}_i^j\|_\infty \leq \|V\|_\infty/2$ when $\mathbf{y}_i^j \in V$.

By taking the logarithm of the above equation, we obtain

$$\begin{aligned} \log p(\mathbf{x}|\mathbf{Y}) &\geq \mathcal{C}_2 - n\frac{\|V\|_\infty}{2\sigma} + \sum_i \log(\sum_{\hat{j}} p(\mathbf{y}_i^j)) \\ &= \mathcal{L}(V), \end{aligned} \tag{13}$$

where $\mathcal{C}_2$ is a constant, n is the number of MPEs, then $\mathcal{L}(V)$ is the lower bound of $\log p(\mathbf{x}|\mathbf{Y})$. Searching for the optimal $\mathbf{x}^*$ now amounts to finding V^* with the largest $\mathcal{L}(V^*)$, *i.e.* maximizing the lower bound of $\log p(\mathbf{x}|\mathbf{Y})$. When σ is very small (the extreme case is $\sigma \to 0$, and it is degenerated to the discrete case), the $\frac{\|V\|_\infty}{2\sigma}$ term is dominant. Under this condition, maximizing $\mathcal{L}(V)$ is equivalent to minimizing $\|V\|_\infty$, which is equivalent to our discrete solution.

We maximize Eq. 13 by using a similar branch and bound method as in Algorithm 1. To further speed up the branch and bound process, we derive the lower and upper bounds of $\mathcal{L}(V) - \mathcal{C}_2$, respectively:

$$f^+(V) = \sum_i \log(\sum_{\hat{j}} p(\mathbf{y}_i^j))$$

$$f^-(V) = -n\frac{\|V\|_\infty}{2\sigma} + \sum_i \log(\sum_{\hat{j}} p(\mathbf{y}_i^j)).$$

Here $f^+(V)$ is obtained by putting all of the modes inside V at the center of the orthotope V, and $f^-(V)$ is obtained by putting all the modes inside V at the boundary of the orthotope V. Neglecting the constant terms, $f^+(V)$ and $f^-(V)$ provide upper and lower bounds of $\mathcal{L}(V)$, respectively. The branch and bound technique can be further accelerated by using these upper and lower bounds for more efficient pruning.

5. Experiments

We evaluate our new MPE fusion methods in two tracking scenarios: one is to track articulated objects (testing the max fusion), and the other is to track occluded objects (testing the median fusion).

5.1. Tracking Articulated Objects

To track an articulated object, the object is decomposed into several parts, and each part is tracked by an individual part-tracker, as explained in Figure 5. As the part-trackers are connected and influence each other, the final tracking result is obtained by fusing the results from the set of part-trackers.

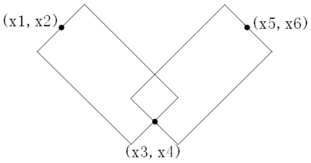

Figure 5. Example of a two-part articulated body. The CE is $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6]$. MPE 1 provides the estimation of $[\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4]$, and MPE 2 provides the estimation of $[\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6]$.

The flowchart of our tracking is shown in Figure 6. In our experiments, each part-tracker is manually initialized by a fixed size rectangle which covers one part of the object. During the tracking process, each part-tracker randomly samples image patches in its neighborhood regions. These image patches are of the same size as the initialized rectangle. To track these patches, we check if their appearances are similar to their initialized appearance, *e.g.* using the sum-of-squared-differences (SSD) measurement. When the similarity score is higher than a predefined threshold, we treat the corresponding coordinates of the matched location as one mode in the partial estimate of the location of the part-tracker. The collection of all such modes gives the MPE of the corresponding tracker. To tolerate appearance variations, we use a less rigid matching criterion that leads to many false positive modes (Figure 7). In our approach, we resample about 1,000 patches and obtain about 20∼200 modes for each part-tracker.

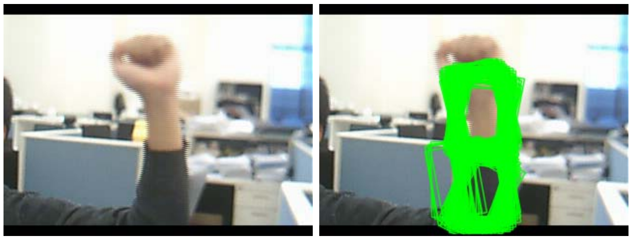

Figure 7. The left figure is the input frame, and the right figure shows the modes of two part-trackers. Two part-trackers handle the upper and lower arms, respectively. Each part-tracker generates 20∼200 modes (shown in green rectangles) in one frame.

After the MPEs are obtained, we apply our new fusion method and compare its performance to RANSAC. In the RANSAC approach, we iterate 1,000,000 ∼ 5,000,000 times. In each iteration, we randomly select one mode from each MPE to obtain the CE by averaging the selections. Then we choose the closest mode to CE from each MPE and calculate the SSD. The experiment settings are the same in our approach as in RANSAC, except for the fusion step. By increasing the number of iterations in RANSAC, it can give good results on tracking a two-part arm. However, for an articulated object that has more than two parts, the RANSAC method performs poorly even if we increase the number of iterations in RANSAC. The execution time of our algorithm is almost fixed when we increase the number of modes in each part-tracker. This shows the good scalability of our algorithm to the number of modes from each MPE.

We test our new fusion method on tracking different articulated objects, and some sample results are shown in Figure 8 and 9. Figure 8 shows the experiment results of tracking articulated objects. By optimizing globally, our algorithm can keep tracking the structure of the articulated object. From top to bottom, the articulated objects have 3, 4, 5 parts, respectively. Figure 9 shows the comparison between our algorithm and RANSAC. The 1st and 2nd row of the figure shows the tracking results of a two-part articulated arm, and both our algorithm and RANSAC can provide good results. The 3rd and 4th row of the figure shows the tracking results of a three-part articulated finger, our algorithm is able to track the finger successfully, but RANSAC fails to give correct results: it begins to drift after several frames. From further experiments, we observe that our fusion method is able to successfully find a global optimum, and outperforms RANSAC. In general, we observe that the more parts we have in the fusion, the better our new fusion method achieves comparing with RANSAC.

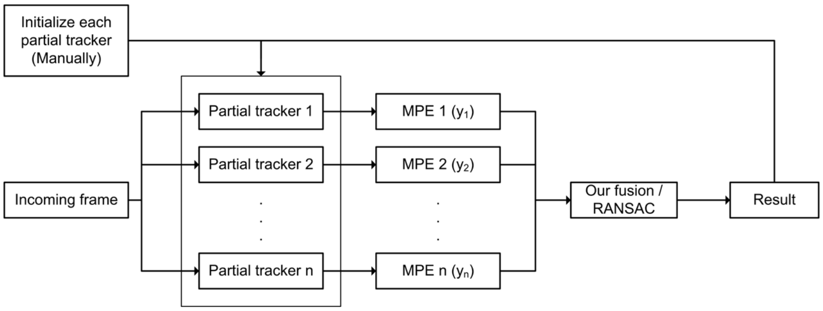

Figure 6. The flow chart of our tracking experiments.

5.2. Tracking Occluded Objects

We evaluate the median fusion in tracking an occluded object. The experiment setting keeps the same as that in the articulated objects tracking. The only difference is that each part-tracker follows a certain part of the object, rather than an articulated part. For example, in the face sequence shown in Figure 10, the face is modeled by eight overlapping parts. Instead of tracking the whole face, we track the eight overlapping face patches. Although every individual part-tracker tracks one of the eight parts and induces many false estimates, the fusion of all the part-trackers leads to a strong tracker which is very robust to partial occlusion. As long as half of the eight patches are visible, the median fusion is able to successfully handle the severe occlusion.

6. Conclusions

Fusing partial estimates from different sources is challenging because of the multimodal nature of the partial estimates: a multimodal objective function can make the optimization process easily trapped in local minima. Generally, it is difficult to obtain the global optimal estimation, especially in a high-dimensional parameter space. By revealing the connection between the probabilistic data fusion and computational geometry, we present a novel approach to the above challenges. We relate the error minimization problem of MPE fusion to a computational geometry problem of finding the minimum-volume orthotope in the parameter space. A branch and bound search algorithm is designed to obtain the global optimal solution. Our proposed new fusion method is scalable w.r.t. the number of estimates and its complexity is almost constant w.r.t. the number of partial estimates. Our proposed algorithm can be applied to a wide variety of applications (*e.g.* articulated objects tracking, occluded objects tracking), where effective information fusion from separate sources is needed.

Acknowledgement

This work was supported in part by National Science Foundation grant IIS-0347877, US Army Research Laboratory and the US Army Research Office under grant ARO W911NF-08-1-0504, and NIH #5 R24 HD050821-03.

References

[1] A. Adam, E. Rivlin, and I. Shimshoni. Robust Fragments-based Tracking using the Integral Histogram. In *CVPR*, 2006.

[2] D. Comaniciu. Nonparametric information fusion for motion estimation. In *CVPR*, volume 1, June 2003.

[3] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[4] B. Han and L. Davis. Probabilistic fusion-based parameter estimation for visual tracking. *CVIU*, 2008.

[5] G. Hinton. Products of experts. In *ICANN*, volume 1, 1999.

[6] G. Hua and Y. Wu. Measurement integration under inconsistency for robust tracking. In *CVPR*, 2006.

[7] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[8] S. Khuller and Y. Matias. A Simple Randomized Sieve Algorithm for the Closest-Pair Problem. *Information and Computation*, 118(1):34–37, 1995.

[9] C. Lampert, M. Blaschko, T. Hofmann, and S. Zurich. Beyond Sliding Windows: Object Localization by Efficient Subwindow Search. In *CVPR*, 2008.

[10] X. Li, Y. Zhu, J. Wang, and C. Han. Optimal linear estimation fusion-part I: unified fusion rules. *IEEE Transactions on Information Theory*, 49(9):2192–2208, 2003.

[11] E. Simoncelli, E. Adelson, and D. Heeger. Probability distributions of optical flow. In *CVPR*, pages 310–315, 1991.

[12] E. Sudderth, E. Ihler, W. Freeman, and A. Willsky. Nonparametric belief propagation. In *CVPR*, pages 605–612, 2002.

[13] J. Uhlmann. Covariance consistency methods for fault-tolerant distributed data fusion. *Information Fusion*, 4(3):201–215, 2003.

[14] J. Xu, J. Yuan, and Y. Wu. Fusion of Multimodal Partial Estimates. Technical report, Department of EECS, Northwestern University, 2009.

[15] J. Yedidia, W. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, pages 239–236, 2003.

[16] J. Yuan, Z. Liu, and Y. Wu. Discriminative Subvolume Search for Efficient Action Detection. In *CVPR*, 2009.

[17] X. Zhou, A. Gupta, and D. Comaniciu. An information fusion framework for robust shape tracking. *PAMI*, 27(1):115–129, 2005.

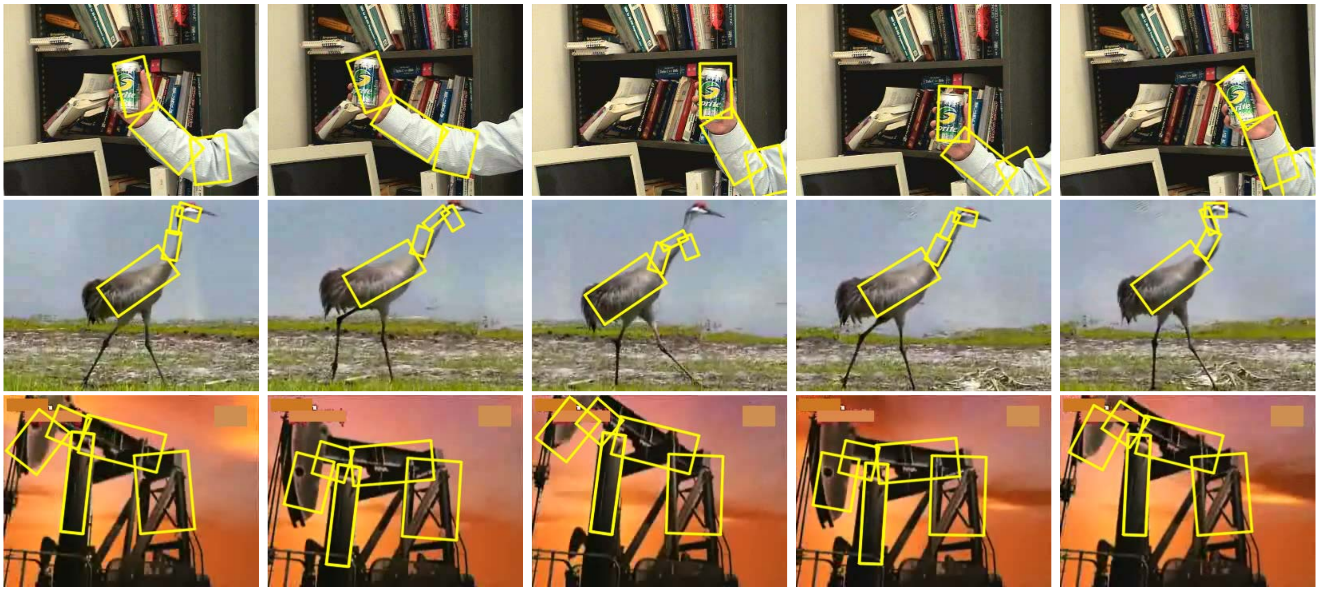

Figure 8. Tracking articulated objects. From top to bottom, the articulated objects are split into 3, 4, 5 parts, respectively.

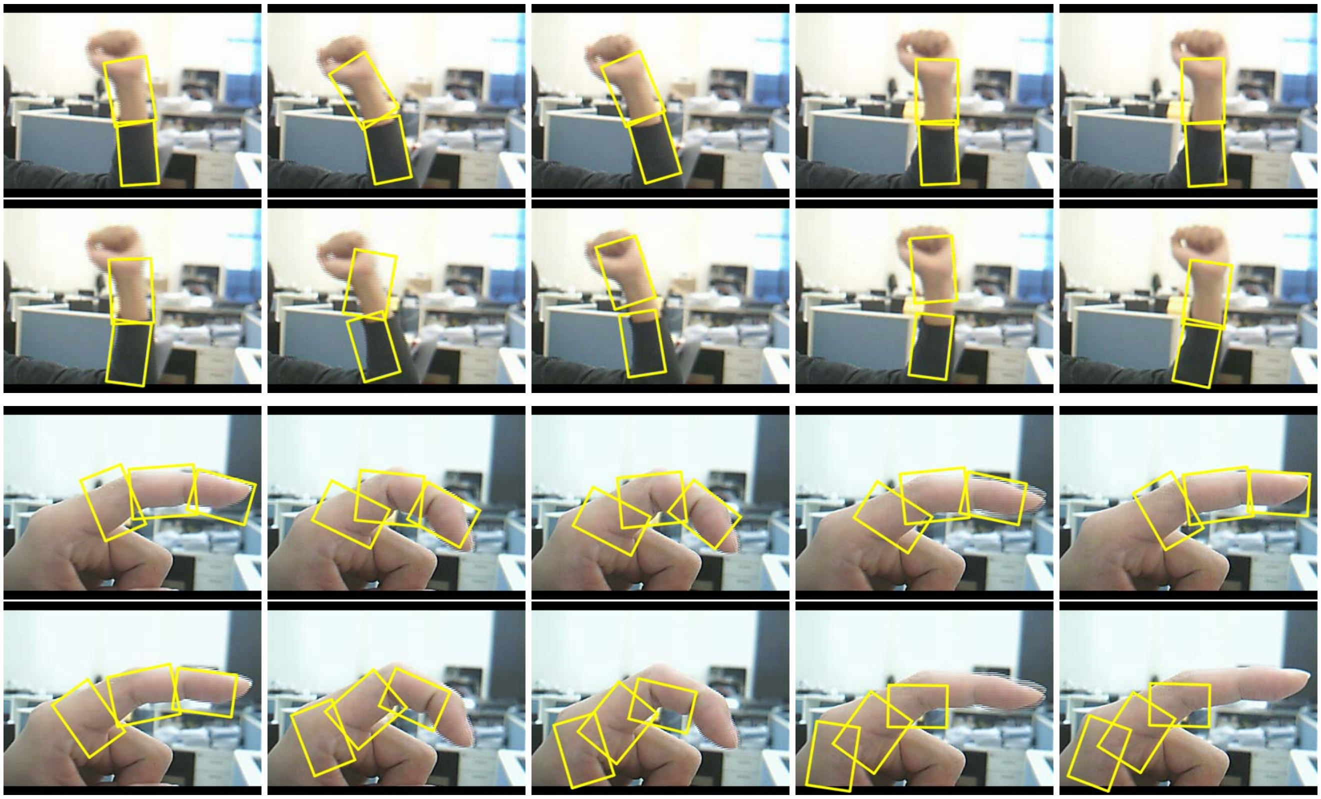

Figure 9. Comparison between our algorithm and RANSAC. Odd rows: our results; even rows: RANSAC results.

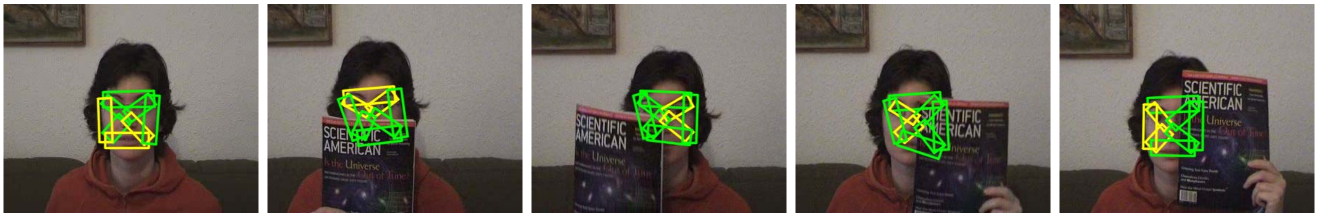

Figure 10. Tracking occluded face (sequence from [1]).