Context-aware Visual Tracking

Ming Yang, *Member, IEEE,* Ying Wu, *Senior Member, IEEE,* and Gang Hua, *Member, IEEE*

Abstract—Enormous uncertainties in unconstrained environments lead to a fundamental dilemma that many tracking algorithms have to face in practice: tracking has to be computationally efficient but verifying whether or not the tracker is following the true target tends to be demanding, especially when the background is cluttered and/or when occlusion occurs. Due to the lack of a good solution to this problem, many existing methods tend to be either effective but computationally intensive by using sophisticated image observation models, or efficient but vulnerable to false alarms. This greatly challenges long-duration robust tracking. This paper presents a novel solution to this dilemma by considering the context of the tracking scene. Specifically, we integrate into the tracking process a set of auxiliary objects that are automatically discovered in the video on the fly by data mining. Auxiliary objects have three properties, at least in a short time interval: (1) persistent co-occurrence with the target; (2) consistent motion correlation to the target; and (3) easy to track. Regarding these auxiliary objects as the context of the target, the collaborative tracking of these auxiliary objects leads to efficient computation as well as strong verification. Our extensive experiments have exhibited exciting performance in very challenging real-world testing cases.

Index Terms—Computer vision, visual object tracking, context-aware, collaborative tracking, data mining, robust fusion, belief inconsistency.

1 Introduction

Robust long-duration visual tracking is demanded by many contemporary applications such as video-based surveillance and vision-based interfaces. One fundamental obstacle in the way is the lack of efficient means for verification, *i.e.*, to determine whether the object being followed by the tracker is really the target. At the extreme, this is in fact a recognition task. Without effective verification, the tracker is likely to drift away gradually, or fail when the target is occluded even for a short period of time. Therefore, although extensive research efforts have been taken, it is still quite difficult in practice to achieve robust and efficient long-duration tracking in unconstrained real-world environments. Most existing methods are in a dilemma: either be fast-but-fallible, or be robust-but-slow.

This dilemma originates from the opposite requirements for the image likelihood models: on one hand, the likelihood model should be simple for efficient motion estimation and tracking; on the other hand, it has to be sophisticated for comprehensive verification of the target. We call them *descriptive* likelihood and *discriminative* likelihood, respectively. In general, descriptive likelihood is based on the descriptive image features that can be easily accessible and specified, *e.g.*, contours [1], [2], colors [3], or even image regions [4], [5], *etc.*. The matching of these image features leads to efficient computation of the descriptive likelihood and thus fast motion estimation (*e.g.*, differential methods such as kernel-based tracking [3], [5], [6]).

However, in practice, many real-world complications such as clutters, illumination and view changes, low image quality, motion blur, and partial occlusions, all may invalidate simple descriptive likelihood models. As a result, good matches of these descriptive features do not necessarily have to correspond to the true target, and background false positive objects may also be good matches. Over the years, there have been two approaches to address this issue: on-line adaptation of the descriptive likelihood models [5], [7]–[9], or using discriminative likelihood models that distinguish the true target from false positives. Without strong verification that provides confident supervision, on-line adaptation is risky and lacks a mechanism to prevent drifting. On the other hand, discriminative likelihood is generally associated with classifiers, *e.g.*, the SVM tracker [10]. These classifiers can be trained off-line or on-line [11], [12]. As learning a classifier has to be based on a large number of training features, it tends to be computationally demanding.

Is there a way to get out of the dilemma so as to have more efficient but still effective verification? In all these existing methods, the dynamic environment is taken for granted as the adverse party for the tracker, as it generates false positives, and most computation has to be spent in separating the true target from the environment. However, the environment can also be advantageous to the tracker if it contains objects that are correlated to the target. For example, if we need to track a face in a crowd, it is almost impossible to learn a discriminative model to distinguish the face of interest from the rest of the crowd. Why do we have to focus our attention only on the target? If the person (with that face) is wearing a quite unique shirt (or a hat), then including the shirt (or the hat) in matching will surely make the tracking much easier and more robust. By the same token, if another face is always accompanying the target face, treating them as a geometric structure and tracking them as a group will be much easier than tracking either of them. It is clear that this makes the verification much

• *Ming Yang and Ying Wu are with Electrical Engineering and Computer Science Department, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208-3118. Email: m-yang4@u.northwestern.edu, yingwu@ece.northwestern.edu. Gang Hua is with Microsoft Research, Redmond, WA 98053. Email: ganghua@microsoft.com.*

Manuscript received June 15, 2007; revised January 3, 2008.

easier as the discriminative model is much simpler. We call this new approach *context-aware tracking* (CAT) as it takes into consideration the context of the target, as shown in Fig. 1.

A target is seldom isolated and independent to the entire scene, therefore there may exist some objects that have short-term or long-term motion correlations to the targets (but are unknown to the tracker beforehand). Thus, taking the advantage of these context information in an efficient way can improve the robustness of the tracker as the spatial context provides additional verification. We represent the context of a target by a set of *auxiliary objects* that are automatically discovered on the fly in an unsupervised fashion by using data mining techniques. A context-aware tracker can discover a set of auxiliary objects and track them simultaneously. Specifically in this paper, auxiliary objects are those that exhibit strong motion correlation to the target. The correlation can be employed to improve tracking and to provide computationally efficient but powerful verification. Intuitively, an auxiliary object should satisfy three properties at least in a short time interval: 1) persist co-occurrence with the target, 2) consistent motion correlation to the target, and 3) easy to track.

In the proposed context-aware tracking, auxiliary objects can be in various forms, *e.g.* solid semantic objects which bear intrinsic relations to the target, or certain image regions that happen to have motion correlation with the target for a short period of time. They may reliably associate to the target for a long duration, or only for a short time interval, or may not exist at all. Thus, it is impossible to determine auxiliary objects off-line in advance, but they have to be discovered on the fly. We resort to data mining techniques for discovering auxiliary objects by learning their co-occurrence associations and estimating affine motion models to the target. Data mining methods originated from text information processing and relational databases [13], and found their uses in extracting video objects [14]–[16]. To the best of our knowledge, this paper presents an original attempt of combining visual tracking and data mining in a collaborative tracking framework.

This new approach has the following advantages. Firstly, it is computationally efficient, because auxiliary objects are easy to track (*e.g.* color regions) and do not incur much computational cost. Secondly, it outputs more accurate tracking results. A context-aware tracker tracks the target and the set of auxiliary objects as a random field in a collaborative manner. It is provably correct that the uncertainty of the motion estimation of the target is reduced. Thirdly, it also provides

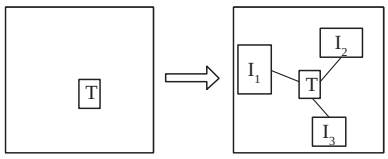


Fig. 1. Illustration of context-aware tracking. T indicates the target and I_k means the spatial context of the target. Traditional tracking methods focus their attention on the target only, while context-aware tracking considers the target and its spatial context within a network.

effective verification, because the learned motion and/or geometric correlations among the target and the auxiliary objects serve as strong cues for verification. Last but not the least, it is intelligent and robust. The context of a target, *i.e.* the auxiliary objects and the motion correlation (*i.e.*, the random field), is automatically discovered on the fly. The robust fusion embedded can handle partial occlusions and even camouflages. Our extensive tests on real-world data give quite exciting performance in dealing with challenging cases including large scale changes, partial occlusions and complicated cluttered backgrounds.

The remainder of this paper is organized as follows. Related work on visual object tracking is reviewed in Sec. 2. The overview of the proposed approach is presented in Sec. 3. The three components of the proposed approach, *i.e.* discovering the auxiliary objects by data mining, collaboratively fusing the tracking results of auxiliary objects and the target, and identifying the outliers, are elaborated in Sec. 4, Sec. 5, and Sec. 6, respectively. Experiments on real-world sequences are reported in Sec. 7. Concluding remarks are in Sec. 8.

2 Related Work

Visual tracking has been an active research topic since the early 1980s and keeps advancing both in theory and practice as the expectations are soaring significantly in real-world applications, *e.g.* video-based security surveillance, medical applications [17], autonomous vehicle [18]. The targets in visual tracking evolve from points in dense optical flow [19]–[22], contours [1], blob regions [3], [23], to more complicated articulated objects [24] and multiple objects [25], [26]. Meanwhile, visual tracking is closely coupled with and greatly benefits from many related tasks, such as background subtraction [27], image/motion segmentation [28], statistical learning [10], [29]. For more comprehensive survey about image features and techniques used in tracking, we refer the readers to [30].

Regardless of the diverse features and targets studied in tracking, essentially as a recursive motion estimation problem, visual tracking mainly involves two fundamental issues: matching and searching. They correspond to target likelihood/observation models that measure the matching between a hypothesis and the target, and the motion estimation schemes that search for the optimal hypothesis. Motion estimation schemes can be differential and based on gradient descent search [3], [5], and be sampling-based such as particle filters [1], [31] or sequential Monte Carlo. The search may incorporate the prior knowledge about target dynamics, *e.g.*, Kalman filters, multiple hypothesis tracking (MHT) [32], [33], or probability data association filter (PDAF) [34], [35].

Target likelihood/observation model is the core in visual tracking which primarily determines tracking accuracy and efficiency. A target can be described by its visual features, based on which a descriptive likelihood model can be constructed. If the features are unique and invariant to the environment changes, tracking is going to be an easy task. However, in the real-world, the environment is unconstrained and presents tremendous variabilities, it is skeptical if the invariant features

determined in advance (thus the descriptive likelihood model) shall still be valid during the run time. Thus, a short term invalidation of the likelihood model, *e.g.*, the target moves out of the field of view or occlusion is present, is very likely to fail the tracker. A more adverse failure situation is that the tracker is following a false positive that also evaluates a large descriptive likelihood.

To deal with this challenge, various approaches have been proposed in the literature. Despite the versatile formulations, in general, they can be categorized into the following three cases: integrating multiple cues, on-line adaptation of descriptive models, and using discriminative models. Taking into consideration of multiple visual cues lead to a rich descriptive model, *e.g.*, geometry and illumination can be combined [5]. The integration can based on simple heuristics [2] or co-inference [36]. On-line adaption of descriptive models changes the parameters of the likelihood model according to the changes of the environment. For example, an appearance model can be adapted based on EM [23] or based on an incremental updating of the basis of the appearance subspace [7], [8].

Since descriptive likelihoods only check the matching of predefined features, a good match is not necessarily be the true target but a false positive. Therefore, another approach is based on using discriminative likelihood models that distinguish the target from the environment. Such discriminative models can be trained off-line in advance, *e.g.*, the SVM tracker [10] that uses the SVM score as the matching criterion. Since the off-line training is to optimize the global and generic discrimination performance, it may not be accurate enough locally. Therefore, on-line adaptation can also be used for discriminative models. For example, this can be done by on-line selection of discriminative color spaces from a fixed set of predefined color spaces to distinguish the target from the background [37], or by selecting Haar features from a large pool [12], or by learning a set of weak classifiers [11], [38].

In contrast to these existing methods, we propose a novel approach to enhancing the observation model by on-line discovery of some auxiliary objects [39] which can help verify the target tracking results. These auxiliary objects with short-term motion correlation to the target can serve as the context of the target. Tracking the target as well as the auxiliary objects in a collaborative way can effectively reduce the uncertainty of the tracking results and deal with large uncertainties of the environments.

3 Overview of Our Approach

The proposed approach, called *context-aware visual tracking*, or *CAT*, has the following three important components:

- **Mining auxiliary objects** (in Sec. 4): the methods of extracting the candidates of auxiliary objects and mining the associations will be discussed. For auxiliary object candidates, multibody grouping is employed to discover the potential multibody structure from motion and to estimate the affine motion models through subspace analysis. This step not only identifies a set of auxiliary objects, but also learns a random field among them;
- **Collaborative tracking** (in Sec. 5): both the target and the set of auxiliary objects need to be tracked in CAT. Because they are not independent, the tracking is formulated based on a random field and is achieved efficiently by the collaborations among all the individual trackers in the network where an individual tracker influences other trackers as well as receiving influence from others;
- **Robust fusion** (in Sec: 6): for an individual tracker, there may exist inconsistency among the influences it receives and its own image measurements. Handling inconsistency is fundamental and critical to fuse auxiliary object trackers and the target tracker.

The entire procedure of CAT algorithm is summarized in Fig. 2. The details of each component will be explained in the following sections.

4 Mining Auxiliary Objects

4.1 Auxiliary objects

Auxiliary objects (AOs) are the spatial context that can help the target tracker. We abuse a little bit the term "object". In fact, it is not necessary for an AO to be a semantic object. In the tracking scenario, it refers to an informative image region or an image feature that satisfies the following three properties:

1) frequent co-occurrence with the target;
2) consistent motion correlation to the target;
3) suitable for tracking.

Although this definition may cover a large variety of image regions or features, not all of them are appropriate for balancing the complexity and generality. Since the prior knowledge about the target and the environments are in general not accessible, it is preferable to choose simple, generic and low-level auxiliary objects, such as image regions or feature points. Feature points are geometrically significant and provide the most localized information. There are some outstanding work on invariant feature points, *e.g.* [40]–[43]. Although feature points may be salient and therefore suitable for object recognition, they are in general prone to occlusion, lighting and local geometry changes. Thus they are not always stable and reliable in video. In addition, extracting invariant features needs a good amount of computation, which makes it hard to achieve real-time performance. Therefore, although the tracking of feature points can be quite efficient, we generally do not use feature points as auxiliary objects.

Instead, we choose to use significant image regions. Different from localized image feature points, image regions reflect the visual property of a neighborhood, and they tolerate more occlusions and local geometry changes. More importantly, image regions, if selected properly, can be reliably and efficiently tracked, for example, by the mean-shift algorithm [3]. Although texture regions may have invariants and can be very significant, our current implementation does not use them because it takes more computation to spot them than color regions. Therefore, our current treatment for data mining is to discover a set of color regions that are temporally stable and spatially correlated to the target in a video sequence in an unsupervised way.

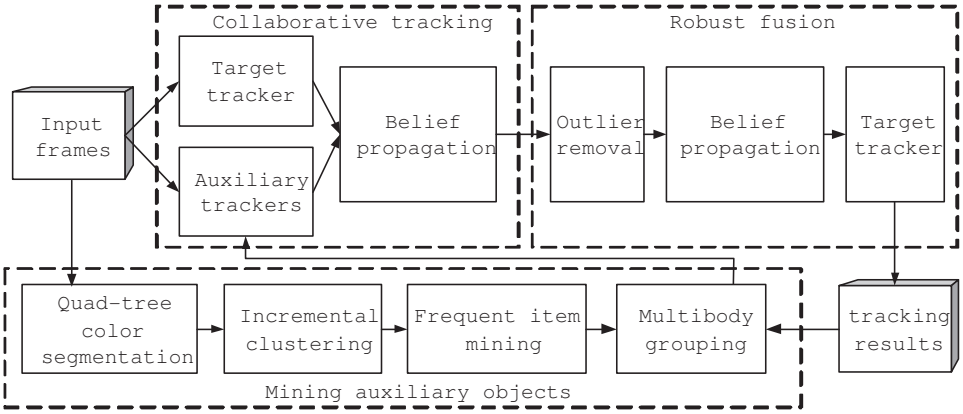


Fig. 2. Block diagram of the CAT algorithm. The sub-modules of auxiliary object mining, collaborative tracking, and robust fusion are enclosed in dash rectangles.

4.2 Item candidate generation

To follow data mining's conventions and make our discussion clear, we define the following terms for our video data mining task.

Definition 1: We denote an *item candidate* by s which is a particular image feature obtained by low-level image processing; an *item* by I which is a quantized item candidate in a *vocabulary* $\mathcal{V} = \{I_1, \ldots, I_N\}$ which is learned by clustering all item candidates; an *itemset* by $\mathbf{I} \subset \mathcal{V}$, set of items; and a transaction by τ, the itemset within a neighborhood R.

In our implementation, an item candidate is a rough color segment with its motion parameters, and an item is defined by $I = \{H(I), \mathbf{x}_I\}$, where $H(I)$ is the average color histogram of the item and $\mathbf{x}_I$ is the motion parameters and respective covariances. The set of candidate AOs, denoted by F, is a subset of $\mathcal{V}$, which are frequently co-occurrent with the target. The candidate AOs that have strong motion correlations to the target are identified as auxiliary objects.

The item candidates s, *i.e.*, the color segments in our case, are the inputs for mining. In the tracking scenario, efficient segmentation is more preferred than a delicate but expensive one since exact boundaries of the segments are not necessary for mining and tracking. In our current implementation, we employ the classical split-merge quad-tree color segmentation [44]. The image is recursively split into the smallest possible homogenous color regions, and then the adjacent regions with similar appearances are merged gradually. The most prominent advantage of this method is computational efficiency. Some segments are not appropriate for tracking, so we employ some heuristics to prune them, *e.g.* segments that are too large (the area over $1/2$ of the entire image) or too small (the area less than 64 pixels), and concave segments (the area less than $1/2$ of the bounding box) are excluded. These kinds of item candidates are suitable for tracking. Fig. 3 shows some typical segmentation results.

4.3 Frequent item mining

Candidate auxiliary objects are the items that are frequently co-occurrent with the target. To build the vocabulary $\mathcal{V}$ so as to construct the transactions for mining, we need to quantize the

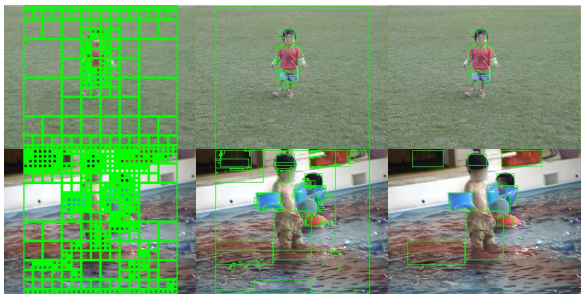


Fig. 3. Illustration of the quad-tree color segmentation. (left) input frame, (middle) over-segmentation, (right) pruned segmentation.

item candidates. In conventional mining applications, usually item candidates can be collected and quantized off-line by k-means or kNN clustering methods. But in this tracking scenario, we have to do this in an incremental way. The procedure is the following. The color segments in each incoming frame are matched to the items in current vocabulary by the Bhattacharyya coefficient [3] of the histograms of the segments as the similarity measurement. Then, each color segment (*i.e.* item candidate) can be quantized and given a label, *e.*g. I_A to I_G are items as shown in Fig. 4. Afterwards, for each item, we form a transaction that consists of the item itself and the items within its neighborhood. There are different choices of the neighborhood. For example, we can use the item itself (*i.e.* use a 0 neighbor). The items inside the region of interest in each frame construct a transaction τ, and a transaction database is built based on M consecutive frames.

Given the transaction database, the items which have a high co-occurrent frequency will be chosen as candidate auxiliary objects. Since the mining is performed online, we need to take into account the importance of the historical images. We maintain an M-frame sliding window and count the item frequency $f(I_n) = \sum_{i=t-M+1}^{t} \beta^{t-i} B_i(I_n)$ with the forgetting factor $\beta = 0.9$ where $B_i(I_n)$ is a binary function and 1 indicates I_n appears in frame i. If image segmentation does not end up with too many small segments, the frequent items are good enough for identifying candidate auxiliary object. If the

segmentation tends to over-segment and produces too many small segments, we cannot use 0 neighbor for constructing transactions, but use the nearby items to form transactions to identify co-occurrent patterns that merge the adjacent small segments. This is another reason that it is fine for image segmentation step to be imperfect. As illustrated in Fig. 4, though there are quite many color segments in each frame, by counting their co-occurrent frequencies, only $F = \{I_A, I_B\}$ are identified as frequent items, *i.e.* candidates of auxiliary objects. The rest of the problem is to determine whether a candidate really bears a motion correlation to the target. The issue will be discussed next.

4.4 Mining by subspace analysis

Finding the frequent items only spots the candidate auxiliary objects that are frequently co-occurrent with the target, but they do not necessarily exhibit strong motion correlations to the target. For example, in Fig. 4, I_B is less correlated to the target T than I_A does. We need to check if these candidates satisfy the motion correlation requirement of an auxiliary object. For each candidate, we can initialize a mean-shift tracker to find its correspondences in the successive image frames. If this tracker loses track for 4 frames in a row, we assert that this candidate is not suitable for tracking and remove it. Otherwise, we can form the motion trajectories over the frames for a set of candidate auxiliary objects. Then, we employ a noise subspace analysis method to discover the potential multibody structure from motion and estimate the affine motion models between the object pairs.

The motion correlation between two moving objects can be very complicated and non-linear, but generally linear motion models can be used as a good approximation. We extend the simple translational model in [39] to a more general affine motion model. When the points on two objects have affine motion relation, they must reside in a linear subspace. Thus, identifying this subspace will lead to the estimation of the affine motion model.

At time t, one candidate auxiliary object $I_O \in F$ is represented as $\mathbf{x}_t = \{u_t^x, v_t^x\}^\top$ and $\{s_t^u, s_t^v\}$ where (u_t^x, v_t^x) are the coordinates of the center of I_O and s_t^u and s_t^v are the scales, respectively. Similarly the target T can be represented as $\mathbf{y}_t = \{u_t^y, v_t^y\}^\top$ and $\{s_t^u, s_t^v\}$. If I_O and T co-occur and have stable motion correlation, then I_O can be claimed as an auxiliary object. So the goal is to evaluate whether I_O and T have strong motion correlation in time window $[t-M+1, t]$ given the trajectories of $\mathbf{y}_t$ and $\mathbf{x}_t$ within this time window.

Assume an affine motion model between candidate auxiliary object I_O and the target T for the period of frame $t-M+1$ to frame t, which is specified by a 2×2 matrix $\mathbf{A}_t$ and a translation vector $\mathbf{b}_t = \{u_t^b, v_t^b\}^\top$, as

$$\mathbf{y}_t = \mathbf{A}_t \mathbf{x}_t + \mathbf{b}_t. \tag{1}$$

Subtract the mean $\bar{\mathbf{y}}_t$ of $\mathbf{y}_t$ and $\bar{\mathbf{x}}_t$ of $\mathbf{x}_t$ in the time window $[t-M+1, t]$ and take the noise into consideration, the relation between I_O and T can be expressed with $\tilde{\mathbf{y}}_t = \mathbf{y}_t - \bar{\mathbf{y}}_t$ and $\tilde{\mathbf{x}}_t = \mathbf{x}_t - \bar{\mathbf{x}}_t$, as

$$\tilde{\mathbf{y}}_t = \mathbf{A}_t \tilde{\mathbf{x}}_t + \mathbf{n}, \tag{2}$$

where $\mathbf{n}$ is a zero mean white noise with $E[\mathbf{n}\mathbf{n}^\top] = \sigma^2 \mathbf{I}$.

If we stack $\tilde{\mathbf{y}}_t$ and $\tilde{\mathbf{x}}_t$, the covariance matrix $\mathbf{C}$ can be expressed as

$$\mathbf{C} = E[\begin{pmatrix} \tilde{\mathbf{y}}_t \\ \tilde{\mathbf{x}}_t \end{pmatrix} (\tilde{\mathbf{y}}_t^\top, \tilde{\mathbf{x}}_t^\top)]. \tag{3}$$

It is clear that $rank(\mathbf{C}) \leq 2$ if there is no noise (*i.e.* $\mathbf{n} = 0$). This rank deficiency property is important in detecting the subspace due to motion correlation. In reality, because $\mathbf{n} \neq 0$, $\mathbf{C}$ is likely to have a full rank. Since the noise is additive, it is easy to prove that the 4D space spanned by $\left(\tilde{\mathbf{y}}_t^\top, \tilde{\mathbf{x}}_t^\top\right)$ is a direct sum of a signal subspace and a noise subspace. The signal subspace is up to rank 2 and corresponds to the large eigenvalues of $\mathbf{C}$, and the noise subspace corresponds to the smallest eigenvalues (*i.e.* σ). Therefore, we can check and threshold the eigenvalues to identify those subspaces.

Denote the estimated covariance matrix by $\hat{\mathbf{C}}$ and the covariance matrix of $\tilde{\mathbf{x}}$ by $\hat{\mathbf{C}}^x$, and we have

$$\hat{\mathbf{C}} = \sum_{i=0}^{M-1} \begin{pmatrix} \tilde{\mathbf{y}}_{t-i} \\ \tilde{\mathbf{x}}_{t-i} \end{pmatrix} (\tilde{\mathbf{y}}_{t-i}^\top, \tilde{\mathbf{x}}_{t-i}^\top) = \begin{pmatrix} \mathbf{A}_t \hat{\mathbf{C}}^x \mathbf{A}_t^\top + \sigma^2 & \mathbf{A}_t \hat{\mathbf{C}}^x \\ \hat{\mathbf{C}}^x \mathbf{A}_t^\top & \hat{\mathbf{C}}^x \end{pmatrix}. \tag{4}$$

Performing eigenvalue decomposition on $\hat{\mathbf{C}}$,

$$\hat{\mathbf{C}} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}, \tag{5}$$

we obtain the sorted eigenvalues $\{\lambda_1, \cdots, \lambda_4\}$ and orthonormal basis $\mathbf{Q}$. If there are more than 2 eigenvalues $\lambda_j^2 \gg \sigma^2$, this candidate is not an auxiliary object since its motion and the target's are not in one subspace.

$$\#\text{ of } \{\lambda_j^2 \gg \sigma^2\} \begin{cases} > 2, & \text{the candidate is not an AO} \\ <= 2, & \text{otherwise} \end{cases}. \tag{6}$$

If the candidate is an auxiliary object, we can estimate its affine matrix $\mathbf{A}_t$ with the property that the noise subspace is orthogonal to the signal subspace. The last two eigenvectors correspond to the noise subspace of $\hat{\mathbf{C}}$ are denoted as

$$\begin{pmatrix} q_{31} & q_{41} \\ q_{32} & q_{42} \\ q_{33} & q_{43} \\ q_{34} & q_{44} \end{pmatrix},$$

which are orthogonal to arbitrary vector $(\tilde{\mathbf{x}}_t^\top \mathbf{A}_t^\top, \tilde{\mathbf{x}}_t^\top)$ in the signal subspace. Substitute them back to $\hat{\mathbf{C}}$, the 2×2 matrix $\mathbf{A}_t$ can be solved by

$$\mathbf{A}_t^\top \begin{pmatrix} q_{31} & q_{41} \\ q_{32} & q_{42} \end{pmatrix} + \begin{pmatrix} q_{33} & q_{43} \\ q_{34} & q_{44} \end{pmatrix} = 0. \tag{7}$$

Then, the translation vector $\mathbf{b}_t$ is obtained with $\bar{\mathbf{y}}_t$, $\bar{\mathbf{x}}_t$, and $\mathbf{A}_t$. This method gives an effective detection of auxiliary objects and efficient estimation of their affine motion models.

Such a mining process is meaningful, because it has learned a random field. We denote the motion of the target T by $\mathbf{y}$ and those of the auxiliary objects by $\mathbf{x}_k, k = 1, \ldots, K$, where K is the number of auxiliary objects. They constitute a random field. The pair-wise potentials $\psi_{k0}(\mathbf{x}_k, \mathbf{y})$ are actually learned as a by-product of this mining process, as

$$\psi_{k0}(\mathbf{x}_k, \mathbf{y}) \propto e^{-\frac{(\mathbf{y}-\mathbf{A}_k\mathbf{x}_k-\mathbf{b}_k)^\top(\mathbf{y}-\mathbf{A}_k\mathbf{x}_k-\mathbf{b}_k)}{2\sigma^2}}, \tag{8}$$

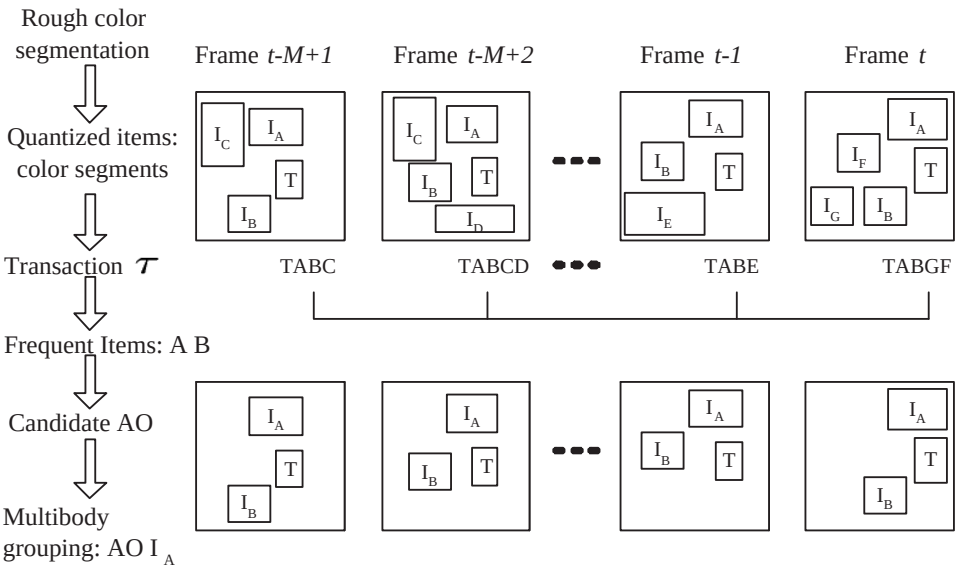


Fig. 4. Illustration of mining auxiliary objects. The target is denoted as T and I_A to I_G represent the items (*i.e.* the color segments). I_A and I_B are selected as candidate auxiliary objects as they are frequently co-occurrent with the target. I_A is identified as one auxiliary object by multibody grouping since it has strong motion correlation to T.

where σ^2 is derived from the small eigenvalues of $\mathbf{C}$ in Eq. 3. In many cases, auxiliary objects share almost the same motion as the target, *e.g.*, the torso and the target head. Therefore, we can use a Gaussian distribution to characterize those potentials. The mean of the Gaussian is given by $\mathbf{A}_k$ and $\mathbf{b}_k$, which is the affine motion model estimated for the kth auxiliary object. Note from now on, the subscript indicates the index of an auxiliary object instead of the time step.

5 COLLABORATIVE TRACKING

It is clear that CAT is not tracking a single target, but a random field. This random field among auxiliary objects and the target is hidden and they need to be inferred from image evidence. We formulate this problem under a Markov network with a special topology, as shown in Fig. 5, where we only assume pair-wise connections between the target $\mathbf{y}$ and the auxiliary object $\mathbf{x}_k$ and there are no connections among auxiliary objects. Each of them is associated with its image evidence $\mathbf{z}_k$. We denote $\mathbf{Z} = \{\mathbf{z}_k, k = 0, \ldots, K\}$, where K is the number of AOs and $\mathbf{z}_0$ is the observation of $\mathbf{y}$ (*i.e.* the target). The core of tracking is to estimate the posteriors $p(\mathbf{y}|\mathbf{Z})$ of the target and $p(\mathbf{x}_k|\mathbf{Z}), k = 1, \ldots, K$, for the auxiliary objects.

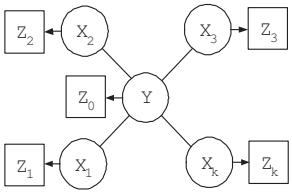


Fig. 5. The star topology of a random field. The hidden motion parameter of the target is denoted as $\mathbf{y}$ with the image observation $\mathbf{z}_0$. The motion parameters of the auxiliary objects are denoted as $\mathbf{x}_k$ with their respective observations $\mathbf{z}_k$.

For such a graph with a star topology, a belief propagation algorithm with 2-step message passing gives the exact estimates of the posteriors. Denote by $p(\mathbf{z}_i|\mathbf{x}_i)$ the local likelihood and by $\phi_k(\mathbf{x}_k)$ the local prior such as the dynamics prediction prior for $\mathbf{x}_k$. Each pair of the target and an auxiliary object $\mathbf{x}_k$ bears a pair-wise potential $\psi_{k0}(\mathbf{x}_k, \mathbf{y})$ learned in the subspace-based mining process, as described in Sec. 4.4. $m_{k0}(\mathbf{y})$ represents the message passed from the kth auxiliary object to the target and $m_{0k}(\mathbf{x}_k)$ is the message from the target to the kth auxiliary object.

At the first iteration step, the target $\mathbf{y}$ receives all the messages m_{k0} from every auxiliary object $\mathbf{x}_k$, then propagates the message back to them at the second iteration. This message passing mechanism implies a collaborative way of tracking. Notice that if the target and the auxiliary objects are independent, their independent motion estimates are $\hat{p}_k(\mathbf{x}_k|\mathbf{Z}) \propto \phi_k(\mathbf{x}_k)p(\mathbf{z}_k|\mathbf{x}_k), k = 1, \ldots, K$. The relation between the true estimates and independent estimates is simply captured by a fixed-point equation of the messages:

$$p(\mathbf{y}|\mathbf{Z}) \propto \hat{p}_0(\mathbf{y}|\mathbf{Z}) \prod_k m_{k0}(\mathbf{y}), \tag{9}$$

$$m_{k0}(\mathbf{y}) = \int_{x_k} \hat{p}_k(\mathbf{x}_k|\mathbf{Z})\psi_{k0}(\mathbf{x}_k, \mathbf{y})d\mathbf{x}_k, \tag{10}$$

$$p(\mathbf{x}_k|\mathbf{Z}) \propto \hat{p}_k(\mathbf{x}_k|\mathbf{Z})m_{0k}(\mathbf{x}_k) \quad k = 1, \ldots, K, \tag{11}$$

$$m_{0k}(\mathbf{x}_k) = \int_y \hat{p}_0(\mathbf{y}|\mathbf{Z}) \prod_{\mathbf{x}_i \setminus \mathbf{x}_k} m_{i0}(\mathbf{y})d\mathbf{y}. \tag{12}$$

This suggests that we can use individual trackers for the target and auxiliary objects. But these set of individual trackers are not independent, as they need to combine their local estimates and the messages from others, and iterate. Such a collaborative mechanism leads to a very efficient solution to tracking the random field. Thus, even if our new approach involves the

tracking of a set of auxiliary objects (*e.g.* by mean-shift), the computation is manageable because of the efficiency of the collaborative way.

Compared with a single tracker for the target, the involvement of auxiliary objects can reduce the uncertainty of the motion estimation of the target and thus make the tracking more confident. We can prove this in a special case when setting both the potential $\psi_{k0}(\mathbf{x}_k, \mathbf{y})$ to be a Gaussian $N(\mu_{k0}, \mathbf{\Sigma}_{k0})$ and the local likelihood $p(\mathbf{z}_k|\mathbf{x}_k)$ to be a Gaussian $N(\hat{\mu}_k, \hat{\mathbf{\Sigma}}_k)$ (we ignore the local prior without losing generality). Under this setting, the closed-form belief propagation gives:

$$\mathbf{\Sigma}_0{}^{-1} = \hat{\mathbf{\Sigma}}_{\mathbf{0}}^{-1} + \sum_{k=1}^{K} (\hat{\mathbf{\Sigma}}_k + \mathbf{\Sigma}_{k0})^{-1}, \tag{13}$$

$$\mu_0 = \mathbf{\Sigma}_0 (\hat{\mathbf{\Sigma}}_{\mathbf{0}}^{-1} \hat{\mu}_0 + \sum_{k=1}^{K} (\hat{\mathbf{\Sigma}}_k + \mathbf{\Sigma}_{k0})^{-1} (\hat{\mu}_k + \mu_{k0})), \tag{14}$$

where $(\mu_0, \mathbf{\Sigma}_0)$ is the target's posterior when tracking the random field. If we assume the local priors to be Gaussian, this result still holds but now $(\hat{\mu}_k, \hat{\mathbf{\Sigma}}_k)$ refers to the local posterior.

Eq. 13 makes it clear that $\mathbf{\Sigma}_0$ is always less than $\hat{\mathbf{\Sigma}}_0$ since these covariance matrices are positive definite and different motion parameters are uncorrelated. Therefore, the confidence of the collaborative estimate of the target is higher than that produced by a single target tracker.

6 Inconsistency and Robust Fusion

The closed form analysis for the collaborative tracking can be explained in the view of information fusion. When the connection potentials between the target and the auxiliary objects are set to be extremely tight, *i.e.*, the covariance of $\mathbf{\Sigma}_{k0}$ is a zero matrix $\mathbf{0}$, this belief propagation is equivalent to the best linear unbiased estimator (BLUE) for $\mathbf{y}$; if they are extremely loose, *i.e.* $\mathbf{\Sigma}_{k0}$ approaches infinity, it becomes an independent estimation; otherwise, it is similar to covariance intersection [45].

However, there is a hidden assumption for this conclusion, *i.e.*, the estimates from all the sources must be consistent. In simple terms, they must more or less agree with each other. But in reality, this may not be valid, when the estimates from the individual trackers may be completely different or inconsistent for many reasons. If using the above mentioned method to fuse these inconsistent estimates, we may end up with an estimate that is completely wrong but of a very high confidence. Such an adverse estimation makes no sense and should be avoided. It is desirable to have a mechanism to detect the inconsistency and identify outliers for a robust fusion.

In this paper, we define two Gaussian sources are *consistent* if the variance in the compatible function of these two Gaussian sources approaches zero using EM estimation (more rigorous and detailed definition is given in Appendix A). In this sense, we proposed a new theorem to measure the consistency for pair-wise Gaussian sources in Markov network [46]. We employ the following two criteria that are very useful for detecting the pair-wise inconsistency. The proofs are presented in Appendix B.

Theorem 1: Considering two Gaussian sources $N(\mu_1, \mathbf{\Sigma}_1)$ and $N(\mu_2, \mathbf{\Sigma}_2)$, where $\mu_1, \mu_2 \in \mathbb{R}^n$, the two sources are inconsistent if:

$$\frac{1}{n}(\mu_1 - \mu_2)^\top (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^{-1} (\mu_1 - \mu_2) \geq 2 + \sqrt{C_p} + \frac{1}{\sqrt{C_p}}, \tag{15}$$

where C_p is the 2-norm condition number of $\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2$, and they are consistent if:

$$\frac{1}{n}(\mu_1 - \mu_2)^\top (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^{-1} (\mu_1 - \mu_2) < 4. \tag{16}$$

Although these are sufficient conditions in general cases, they are actually also necessary conditions when $n = 1$. These criteria enable simple and quick detection of pair-wise inconsistency. Then, the estimation that is inconsistent with all the others will be regarded as an outlier. The outlier can be the target or the AOs. If the target is an outlier, we assert that the target is experiencing occlusion or drift, and suspend the mining process temporarily. In this case, we can give an estimation of the target purely based on the predictions from the auxiliary objects, and search for the image evidence. If the outlier is an auxiliary object, we simply exclude this auxiliary object from fusion. After excluding the outliers, we perform belief propagation again on the rest of the network and employ the target tracker to locate the target precisely. When the majority are not consistent which means the target estimate can not be verified, a tracking failure is asserted.

7 Experiments

7.1 Experiment settings

We substantialized and implemented the proposed CAT algorithm in a head tracking system, where the head tracker is a contour-based elliptical tracker similar to [2], and the auxiliary trackers are mean-shift trackers. Since a fixed number of edge points along the ellipse are matched, the single head tracker is quite computationally efficient and runs at over 50 frame per second (fps). Although the single head tracker is relatively robust to illumination and view changes, it is vulnerable to cluttered backgrounds, motion blur and occlusions. In our experiments, we compare the proposed CAT algorithm to the single head tracker in a large number of real-world sequences captured in unconstrained environments including both indoor and outdoor scenes. These extensive experiments and exciting results have demonstrated the advantages of the CAT algorithm. Furthermore, we apply the same CAT algorithm to people tracking based on an appearance-based torso tracker to exhibit the applicability of the proposed idea to different types of targets.

The motion parameter $\mathbf{y} = \{u, v, s^u, s^v\}$ to be recovered includes the location (u, v) and the scales s^u and s^v. The color segmentation and the mean-shift tracker work in the normalized R-G color space with 32×32 bins. Without code optimization, our C++ implementation of CAT comfortably runs at around 10 fps on average on a Pentium 3GHz desktop for 320×240 images depending on the number of auxiliary objects discovered.

7.2 Quantitative experiments

For a quantitative evaluation, we manually labelled the ground truth of the sequences `kid in yellow`, `dancing girl` and `birthday kid` for 1200, 1600 and 1460 frames respectively. The evaluation criteria of tracking error are based on the relative position errors between the center of the tracking result and that of the ground truth, and the relative scale normalized by the ground truth scale. Ideally, the position differences should be around 0, and the relative scales 1.

As shown in Fig. 6, Fig. 7 and Fig. 8, the position differences of the results in the CAT are much smaller than that of the single head tracker and the relative scales have much less fluctuations around 1. It demonstrates the advantages of the CAT, *i.e.* reducing the false alarm rate and the estimation covariance. Note that at the end of the sequence `kid in yellow`, the single tracker happens to track the head by chance after the drift. Although the CAT tracker loses track at around frame 1100 for several frames, it is able to recover promptly because of the auxiliary objects.

Some key frames are shown in Fig. 9 [1] The first row shows the results of the single head tracker where the highlighted solid-yellow box indicates the location of the head. The second row is the segmentation and mining results, where each green rectangle indicates an item in the current frame. The numbers in blue at the corner show the item labels of the candidate auxiliary objects. The third row illustrates the fusion results. Each blue box is the estimate of the target from difference sources (*i.e.* the target or the auxiliary objects trackers). The white box indicates that estimate is regarded as an outlier. The dark red box is the final result of the fusion. The corresponding labels of the auxiliary objects are shown at the bottom-right corner. The final tracking results of CAT are shown in the 4-th row as highlighted solid-yellow boxes, and the dash-red boxes are the auxiliary object trackers.

7.3 Occlusion and drift

Fig. 9 samples the results on the sequence `kid in yellow` which is very challenging due to a serious occlusion, target out-of-range and the clutters. When the head moves outside the upper boundary at frame 113, the single head tracker drifts to a false positive in the cluttered background and is unable to recover. In contrast, the CAT tracker asserts the occlusion and keeps tracking correctly. It freezes the head tracker temporarily and re-initializes it based on the predictions provided by the auxiliary objects. When the kid is walking in front of the bush, the background is so cluttered that it causes big troubles to the edge-based tracker. On the other hand, CAT discovers several auxiliary objects, *i.e.* the shirt and short pant, which are quite stable and provide roughly correct estimates of the head location and rescue the head tracker from the drift at frame 736.

7.4 Quick movement and camouflage

As shown in Fig. 10, the sequence `dancing girl` presents quick movements and camouflage. All the girls are similar in terms of their appearances. This is extremely difficult for a single head tracker to work, but CAT comfortably handles such a challenge. During the dancing, CAT gradually discovers the spatial relations between the target (the girl of interest) and the adjacent context *e.g.* other girls' shirts, although such relations are only valid in a short time interval. At frame 757, the single head tracker is trapped by the shoulder of the girl and unable to recover. At frame 758, the CAT tracker identifies this false alarm and pulls back the head tracker with the help of the predictions of the AOs that are close to the true target. At frame 1234, the girl of interest suddenly bows down, CAT detects the tracking failure and resumes tracking quickly. CAT can comfortably track over 1600 frames for this highly dynamic sequence until the target moves outside the left boundary for several seconds.

7.5 Scale and view changes

We show the tracking performance when the target undergoes large scale and view changes and demonstrate the transition of the auxiliary objects in the sequence `kid&dad` (Fig. 11). For the single head tracker, when the scale of the head becomes very small, it drifts to the torso of the kid from frame 69 and fails the tracker. During the first 300 frames, the dad walks with the kid with quite stable motion correlation. This is discovered by CAT and the region of dad's shirt is mined as the auxiliary object to help track the kid's head. When they move close to the camera, the scale and the view change dramatically so that the learned relation between dad's shirt and the kid's head no longer holds. Fortunately, CAT spots that the hat is a good auxiliary object at large scale and guides the tracking. At the end of the video, the head is completely occluded by the hat for several seconds. Although this is impossible to recover, CAT detects and reports the tracking failure, while the single head tracker tends to drift to a false positive without notice.

7.6 Cluttered background

In sequence `birthday kid`, the target head experiences large out-of-plane rotation and the appearances change greatly, as shown in Fig. 12. For the contour tracker, when the rear head is in the dark background, no good observation is available around the head so the contour tracker drifts to the torso and other elliptical regions, and is unable to recover. For the CAT tracker, with the help of the auxiliary objects, the tracker either keeps tracking in the tough situations or recovers from drifting in several frames. Note the auxiliary objects discovered can be some objects with inherent relations with the target, such as the hat and short pant, or just something that happens to have temporary relations, such as the refrigerator or the gift box. This real-world sequence demonstrates the advantages of the auxiliary objects for long-duration tracking.

As shown in Fig. 13 (`swimming boy`), the background is quite cluttered due to the texture of water and other people, which makes the single head tracker hopeless. The single head tracker is easily distracted by the edges in the background and drifts away. On the other hand, CAT discovers the two blue life buoys and the swimming hat and uses them as the auxiliary objects. When the boy jumps towards his mother's arms, CAT

1. All the faces shown in this paper were mosaicked for privacy protection.

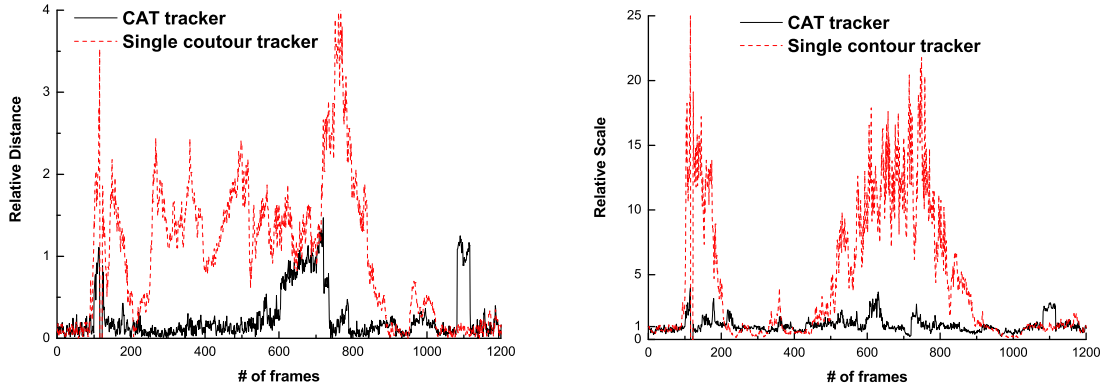


Fig. 6. Quantitative comparison: (left) position errors, (right) scale errors, [`kid in yellow,1200 frames`].

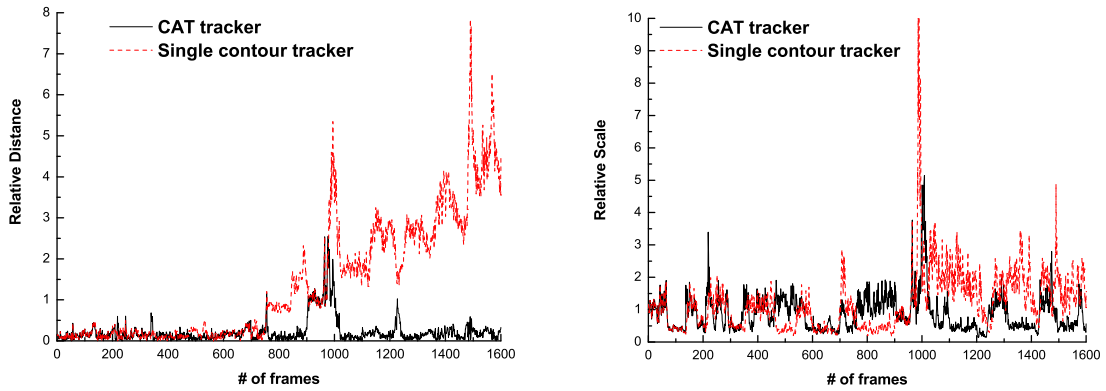


Fig. 7. Quantitative comparison: (left) position errors, (right) scale errors, [`dancing girl,1600 frames`].

uses the life buoys as well as the orange box on the bank to help locate his head accurately, which is difficult for the single head tracker. Note that at the end of this sequence, the kid's head is occluded by his mom's head and CAT switches to the mom. This is reasonable because the auxiliary objects can not differentiate the two heads at the same location.

7.7 More people tracking results

To demonstrate the generalization ability of the proposed method, we apply the context-aware tracking algorithm to people tracking based on an appearance-based torso tracker. As shown in Fig. 14 [47], when the person to track is occluded by his friends around frame 56, the single torso tracker loses the target and drifts away. In contrast, since the other pedestrians serve as the temporary contexts, they can help the CAT tracker keep following the target. In addition, after frame 135 the context information help to prevent the tracker drift to the person next to the target though both persons have very similar appearances. Another example sequence is shown in Fig. 15 where an athlete in a marathon match is tracked with natural lighting changes and view changes present.

7.8 Discussions

As demonstrated in a large number of challenging sequences, there are two primary scenarios when the auxiliary objects greatly help the tracking: 1) some auxiliary objects have persistent relations to the target and present fairly accurate estimates although these relations may not be foreseen; 2) a number of auxiliary objects have transitional relations to the target and the majority of them can give rough correct estimates in a short time interval. In the cases of occlusion or drift, it is not likely that all the auxiliary objects are occluded or all auxiliary trackers lose track at the same time, since the auxiliary objects may not be located in a close vicinity of the target. The mechanism of robust fusion can identify the inconsistency induced by occlusions or drifts. There are some extremely difficult cases, *e.g.* the target is occluded for long time, and CAT fails reasonably because on-line data mining may not be invoked at all. Or only a couple of auxiliary objects discovered and they do not agree with each other about the target motion, which implies insufficient context information to verify the tracking results. At these cases, the advantage of CAT is the ability to detect and report the failure, and leave the system to other means of re-initialization, while the single

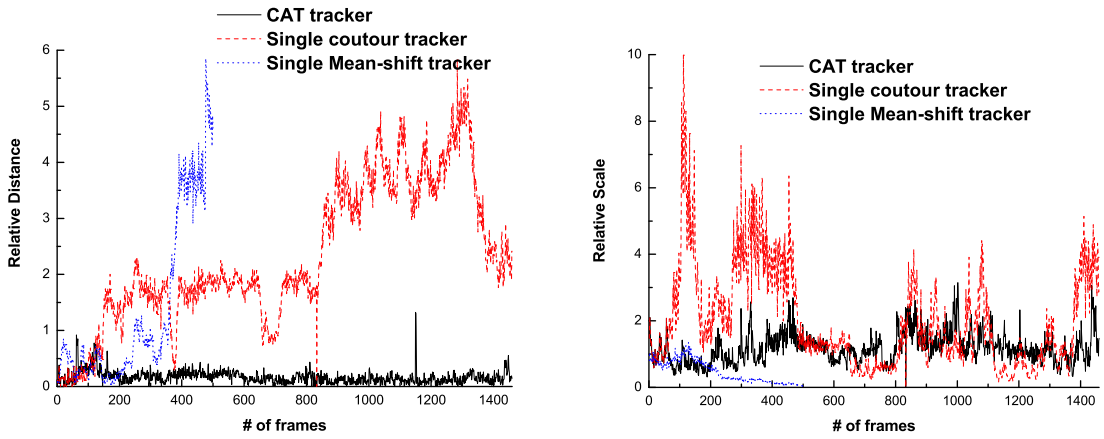


Fig. 8. Quantitative comparison: (left) position errors, (right) scale errors, [`birthday kid,1460 frames`].

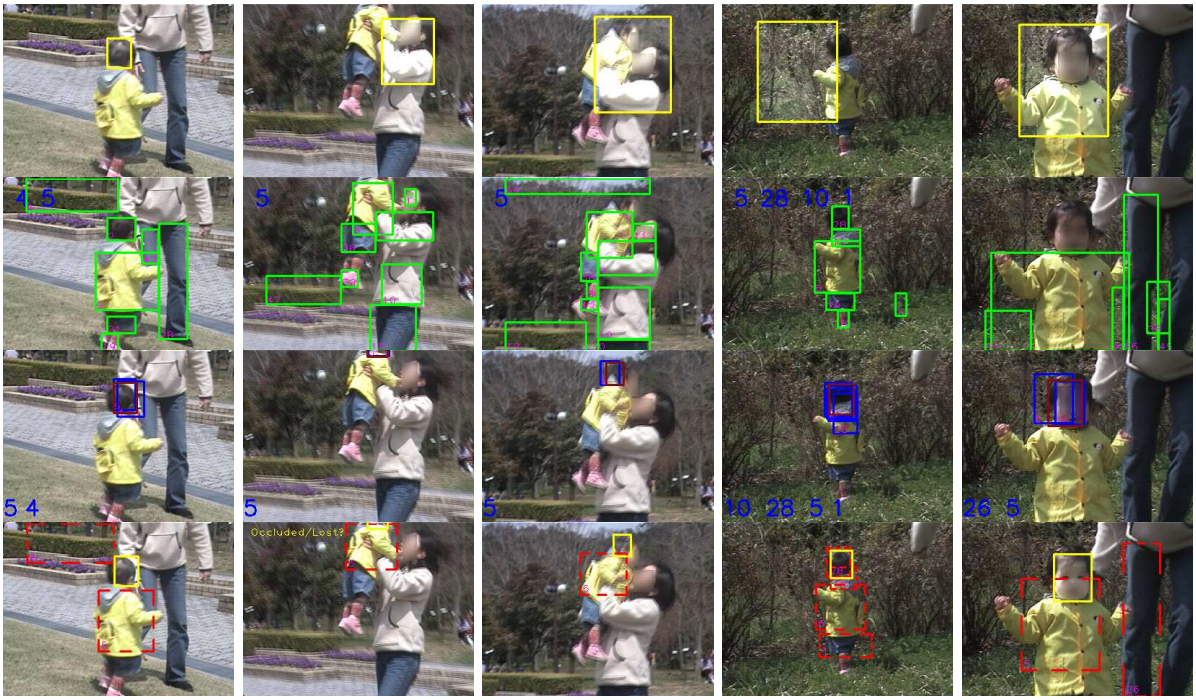


Fig. 9. Frame # 50, 113, 124, 736 and 866 of `kid in yellow,1200 frames`. (1st row) the head tracker, (2nd row) the mining results, (3rd row) the fusion results,(4th row) the CAT tracker.

tracker has no reliable mechanism to report the failure but keep tracking aimlessly and regardlessly. In view of this, the benefit of CAT is pronounced.

8 Conclusions

We have proposed a novel solution to robust long-duration tracking by considering the context of the target. By integrating an unsupervised data mining procedure, a set of auxiliary objects are discovered on the fly which provide extra measurements to the target and reduce the uncertainty of the estimation. In addition, the learned motion correlations among the auxiliary objects and the target serve as a strong cue to verify the tracking results to handle short-term occlusion or tracking lost. The auxiliary objects are automatically discovered without supervision and do not incur much extra computation, which makes the approach generally applicable to a wide spectrum of tracking scenarios.

For future work, we will study the relation between the number of auxiliary objects discovered and the confidence level of the verification. Another important issue to investigate is how to compromise the need for a quicker initial mining procedure within a shorter time window which may find more auxiliary objects and a longer time window which may find less auxiliary objects but with a high reliability.

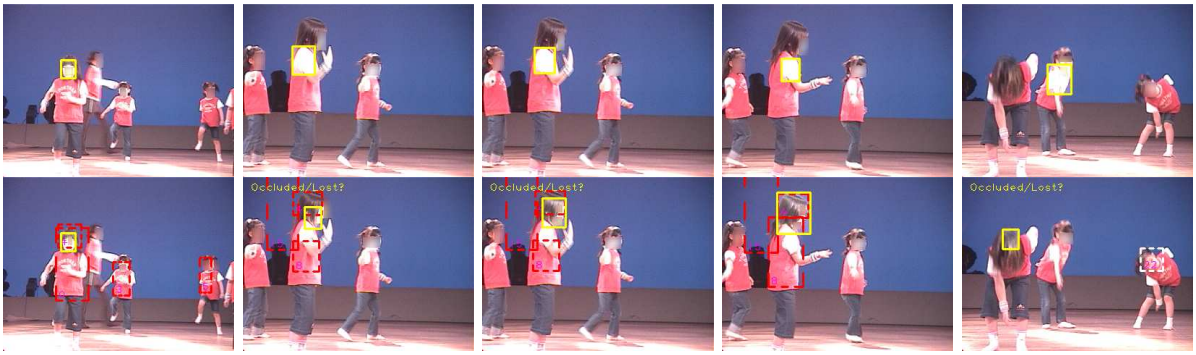


Fig. 10. Frame # 67, 757, 758, 764, and 1234 of `dancing girl,1600 frames`. (top) the head tracker, (bottom) the CAT tracker.

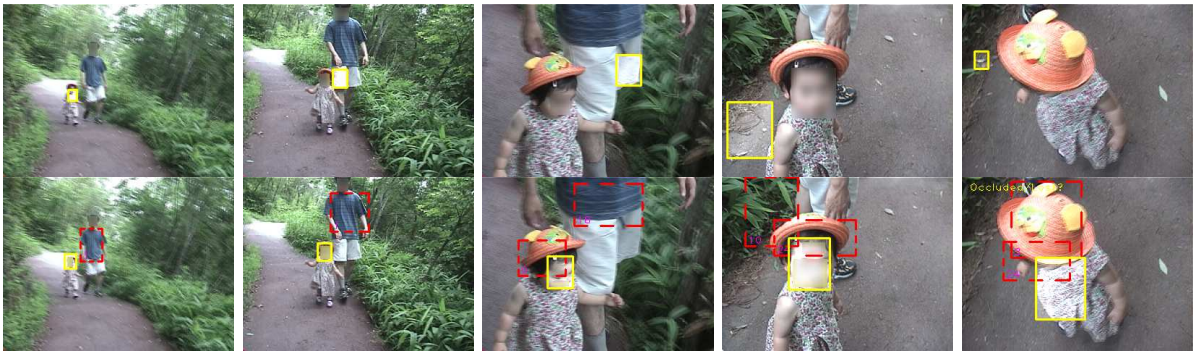


Fig. 11. Frame # 69, 180, 313, 540 and 616 of `kid&dad,617 frames`. (top) the head tracker, (bottom) the CAT tracker.

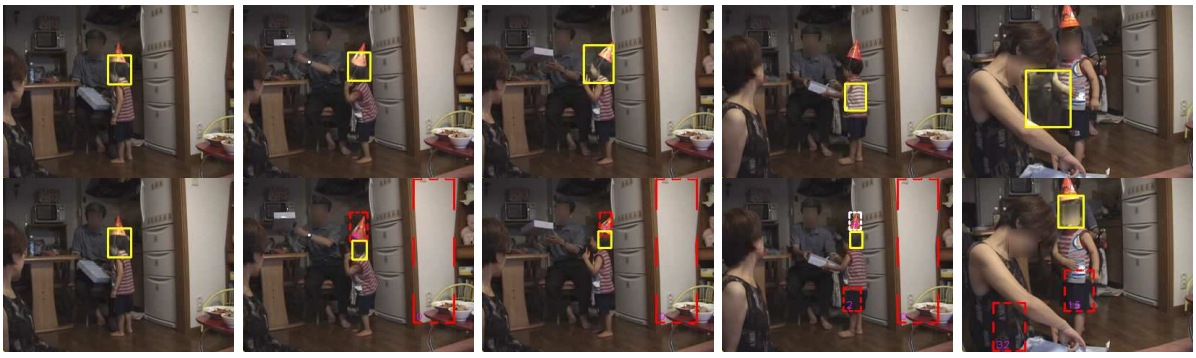


Fig. 12. Frame # 0, 72, 93, 170, and 1455 of `birthday kid,1460 frames`. (top) the head tracker, (bottom) the CAT tracker.

APPENDIX
PROOF OF THEOREM 1

A. Definition of inconsistency in two-node Gaussian Markov network

We consider to define the inconsistency in a two-node Gaussian Markov network, as shown in Fig. 16, where the two observation nodes are Gaussian random vectors $\mathbf{z}_1 \sim N(\mu_1, \Sigma_1)$ and $\mathbf{z}_2 \sim N(\mu_2, \Sigma_2)$ with $\mu_1, \mu_2 \in \mathbb{R}^n$. Therefore, the compatible functions between observation nodes and the hidden nodes are Gaussian, *i.e.*,

$$\phi(\mathbf{x}_i, \mathbf{z}_i) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{\Sigma}_i|}} \mathbf{e}^{-\frac{1}{2}(\mathbf{z}_i - \mathbf{x}_i)^\top \mathbf{\Sigma}_i^{-1} (\mathbf{z}_i - \mathbf{x}_i)}. \quad (17)$$

Assume $\mathbf{x}_1$ can be predicted by a function f of $\mathbf{x}_2$, the compatible or the potential function of $\mathbf{x}_1$ and $\mathbf{x}_2$ can be expressed as a Gaussian

$$\psi(\mathbf{x}_1, \mathbf{x}_2) = \frac{\exp\left\{-\frac{(\mathbf{x}_1 - f(\mathbf{x}_2))^\top (\mathbf{x}_1 - f(\mathbf{x}_2))}{2\sigma_{12}^2}\right\}}{\sqrt{(2\pi)^n \sigma_{12}^n}} \quad (18)$$

$$\doteq \frac{\exp\left\{-\frac{(\mathbf{x}_1 - \mathbf{A}_{12}\mathbf{x}_2 - \mu_{12})^\top (\mathbf{x}_1 - \mathbf{A}_{12}\mathbf{x}_2 - \mu_{12})}{2\sigma_{12}^2}\right\}}{\sqrt{(2\pi)^n \sigma_{12}^n}}, \quad (19)$$

which indicates if $\mathbf{x}_1$ and $f(\mathbf{x}_2)$ can be regarded as being generated from one common model and σ_{12}^2 is the scalar variance. When f is nonlinear, we linearize it by Taylor expansion, *i.e.*, $\mu_{12} = f(\mathbf{0})$ and $\mathbf{A}_{12} = \frac{\partial f_{12}(\mathbf{x}_2)}{\partial \mathbf{x}_2}|_{\mathbf{x}_2 = \mathbf{0}}$ is the $n \times n$ Jacobian. So we only consider the linearized relation of $\mathbf{x}_1$ and $\mathbf{x}_2$ in Eq. 19.

σ_{12}^2 indeed models the uncertainties between the estimate

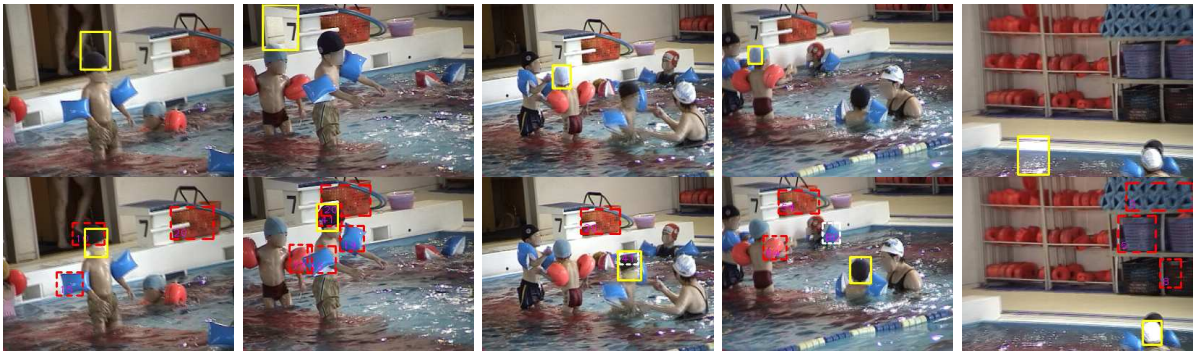


Fig. 13. Frame # 87, 334, 526, 578 and 848 of `swimming boy,900 frames`. (top) the head tracker, (bottom) the CAT tracker.

Fig. 14. Frame # 40, 56, 68, 135, and 425 of `three past shop,425 frames`. (top) the torso tracker, (bottom) the CAT tracker.

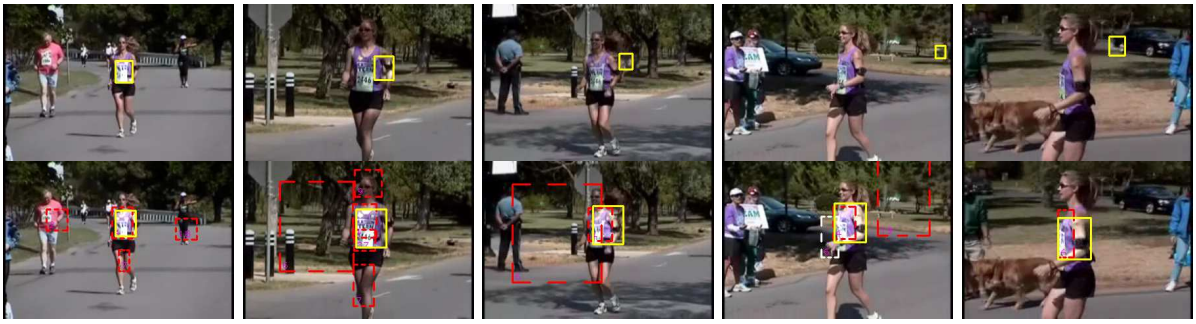


Fig. 15. Frame # 72, 468, 504, 582, and 625 of `marathon, 625 frames`. (top) the torso tracker, (bottom) the CAT tracker.

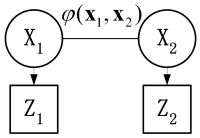


Fig. 16. A two-node Markov network.

$\mathbf{x}_1$ and the neighborhood estimate $\mathbf{A}_{12}\mathbf{x}_2 + \mu_{12}$. Assume $\mathbf{A}_{12}$ and μ_{12} are known, given all the $\{\mathbf{z}_1, \mathbf{z}_2\}$, the estimate of σ_{12}^2 is a natural indicator of whether $\mathbf{x}_1$ and $\mathbf{A}_{12}\mathbf{x}_2 + \mu_{12}$ should be consensus, *i.e.*, if σ_{12}^2 is very small toward zero, then they should be consensus since $\psi(\mathbf{x}_1, \mathbf{x}_2)$ is approaching to an impulse delta function, and vice versa.

The Bayesian MAP inference of $\mathbf{x}_1$ and the ML estimate of σ_{12} can be obtained by the following Bayesian EM algorithm [48], *i.e.*,

$$\begin{aligned} \mathbf{x}_1 &= (\mathbf{\Sigma}_1^{-1} + \frac{1}{\sigma_{12}^2}\mathbf{I})^{-1} \\ &\times (\mathbf{\Sigma}_1^{-1}\mathbf{z}_1 + \frac{1}{\sigma_{12}^2}(\mathbf{A}_{12}\mathbf{x}_2 + \mu_{12})), \end{aligned} \tag{20}$$

$$\sigma_{12}^2 = \frac{1}{n}(\mathbf{x}_1 - \mathbf{A}_{12}\mathbf{x}_2 - \mu_{12})^\top(\mathbf{x}_1 - \mathbf{A}_{12}\mathbf{x}_2 - \mu_{12}). \tag{21}$$

Fixing σ_{12}, the E-Step in Eq. 20 obtains the MAP estimate of $\mathbf{x}_1$ by fixed-point iteration. Fixing $\mathbf{x}_1$ and $\mathbf{x}_2$, the M-Step in Eq. 21 maximizes $p(\mathbf{x}_1, \mathbf{x}_2|\sigma_{12}, \mathbf{z}_1, \mathbf{z}_2)$ *w.r.t.* σ_{12}. Combining the two steps together also constitutes a fixed-point iteration for σ_{12}^2.

We measure the consistency of two observation sources $\mathbf{z}_1$ and $\mathbf{z}_2$ by examining if their estimates $\mathbf{x}_1$ and $\mathbf{x}_2$ are consensus, *i.e.* if $\mathbf{x}_1$ is predictable from $\mathbf{x}_2$ through a linear relation $\mathbf{A}_{12}\mathbf{x}_2 + \mu_{12}$ with small variance σ_{12}^2. Therefore, when

$\mathbf{z}_1$ and $\mathbf{z}_2$ are consistent, the estimate of $\mathbf{x}_1$ and $\mathbf{A}_{12}\mathbf{x}_2+\mu_{12}$ will be consensus, *i.e.*, they will be almost the same. In this case, from Eq. 21, the estimate of σ_{12}^2 will always approach to zero, *i.e.*, zero is the only fixed-point. On the contrary, if they are inconsistent, the estimate of $\mathbf{x}_1$ and $\mathbf{A}_{12}\mathbf{x}_2+\mu_{12}$ may deviate from each other, *i.e.*, the convergent results of σ_{12}^2 may be non-zero. This indicates that there exist non-zero fixed-points for σ_{12}^2. These motivate us to define the inconsistency of two Gaussian sources as follows.

Definition 2: If zero is the only fixed-point for σ_{12}^2 in the Bayesian EM, *i.e.* in Eq. 20 and Eq. 21, $\{\mathbf{z}_1, \mathbf{\Sigma}_1\}$ and $\{\mathbf{z}_2, \mathbf{\Sigma}_2\}$ are *consistent*; if there exist non-zero fixed-points for σ_{12}^2, they are *inconsistent*.

B. Proof of the inconsistency criterion

Given the aforementioned definition of inconsistency for two Gaussian sources in two-node Markov network, we propose a sufficient condition to check the convergent value of σ_{12}^2 as stated in Theorem 1. The basic idea of the proof is to check if Eq. 21 has non-zero solutions. With some manipulations we express Eq. 21 as a function $F(\sigma_{12}^2)$ in Eq. 27. Then, we show if the condition number C_p of $\mathbf{\Sigma}_1+\mathbf{\Sigma}_2$ satisfies Eq. 15 in Theorem 1, there exist two positive numbers $0<k_2<k_1$ such that $F(k_1)<0$ and $F(k_2)>0$, which indicates there is a non-zero solution. If C_p satisfies Eq. 16, $F(\sigma_{12}^2)<0$ for all $\sigma_{12}^2>0$, thus there is no non-zero solution for Eq.21.

Proof: Fixing σ_{12}^2, the fixed-point iteration in Eq. 20 is guaranteed to obtain the exact MAP estimate on the joint posterior Gaussian. For simplification of notation, we denote $\hat{\mathbf{x}}_2=\mathbf{A}_{12}\mathbf{x}_2+\mu_{12}$ and $\hat{\mathbf{z}}_2=\mathbf{A}_{12}\mathbf{z}_2+\mu_{12}$. Define $\mathbf{P}=\mathbf{\Sigma}_1+\mathbf{\Sigma}_2$ and $\mathbf{S}=\mathbf{P}+\sigma_{12}^2\mathbf{I}$. The convergent result in the E-Step in Eq. 20 is the same as,

$$\begin{bmatrix} \mathbf{x}_1 \\ \hat{\mathbf{x}}_2 \end{bmatrix} = \begin{bmatrix} (\sigma_{12}^2\mathbf{I}+\hat{\mathbf{\Sigma}}_2)\mathbf{S}^{-1}\mathbf{z}_1+\mathbf{\Sigma}_1\mathbf{S}^{-1}\hat{\mathbf{z}}_2 \\ \hat{\mathbf{\Sigma}}_2\mathbf{S}^{-1}\mathbf{z}_1+(\sigma_{12}^2\mathbf{I}+\mathbf{\Sigma}_1)\mathbf{S}^{-1}\hat{\mathbf{z}}_2 \end{bmatrix}. \quad (22)$$

Embedding it to the M-Step in Eq. 21, we have

$$\sigma_{12}^2=\frac{1}{n}\sigma_{12}^2\sigma_{12}^2(\mathbf{z}_1-\hat{\mathbf{z}}_2)^\top\mathbf{S}^{-1}\mathbf{S}^{-1}(\mathbf{z}_1-\hat{\mathbf{z}}_2). \quad (23)$$

To prove Theorem 1, since zero is a solution of σ_{12}^2 for Eq. 23, we only need to analyze the existence of non-zero solutions of σ_{12}^2 for

$$\frac{1}{n}\sigma_{12}^2(\mathbf{z}_1-\hat{\mathbf{z}}_2)^\top\mathbf{S}^{-1}\mathbf{S}^{-1}(\mathbf{z}_1-\hat{\mathbf{z}}_2)-1=0. \quad (24)$$

$\mathbf{P}$ is the sum of two covariance matrices so it is *real positive definite*, thus there exists an orthonormal matrix $\mathbf{Q}$ such that $\mathbf{P}=\mathbf{Q}\mathbf{D}_p\mathbf{Q}^\top$, where

$$\mathbf{D}_p=diag[\sigma_1^2,\sigma_2^2,\ldots,\sigma_n^2]$$

is the eigen-matrix with $\sigma_1^2\geq\sigma_2^2\geq\ldots\geq\sigma_n^2>0$ and $C_p=\frac{\sigma_1^2}{\sigma_n^2}$. Then we have $\mathbf{S}=\mathbf{Q}\mathbf{D}_s\mathbf{Q}^\top$, where

$$\mathbf{D}_s=diag[\sigma_1^2+\sigma_{12}^2,\sigma_2^2+\sigma_{12}^2,\ldots,\sigma_n^2+\sigma_{12}^2].$$

Furthermore, $\mathbf{S}^{-1}=\mathbf{Q}^\top\mathbf{D}_s^{-1}\mathbf{Q}$ where

$$\mathbf{D}_s^{-1}=diag[\frac{1}{\sigma_1^2+\sigma_{12}^2},\frac{1}{\sigma_2^2+\sigma_{12}^2},\ldots,\frac{1}{\sigma_n^2+\sigma_{12}^2}].$$

We also denote $\tilde{\mathbf{z}}=\mathbf{Q}(\mathbf{z}_1-\hat{\mathbf{z}}_2)=[\tilde{z}_1,\tilde{z}_2,\ldots,\tilde{z}_n]^\top$. Then, we can simplify the expressions in Eq. 24 and Eq. 15 in Theorem 1 (Sec. 6) as,

$$\frac{1}{n}\sigma_{12}^2(\mathbf{z}_1-\hat{\mathbf{z}}_2)^\top\mathbf{S}^{-2}(\mathbf{z}_1-\hat{\mathbf{z}}_2)=\frac{1}{n}\sum_{i=1}^{n}\frac{\sigma_{12}^2\tilde{z}_i^2}{(\sigma_i^2+\sigma_{12}^2)^2}, \quad (25)$$

$$\frac{1}{n}(\mathbf{z}_1-\hat{\mathbf{z}}_2)^\top\mathbf{P}^{-1}(\mathbf{z}_1-\hat{\mathbf{z}}_2)=\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{z}_i^2}{\sigma_i^2}. \quad (26)$$

From Eq. 25, we express Eq. 24 as a function $F(\cdot)$ of σ_{12}^2 and only need to analyze the solution of σ_{12}^2 for

$$F(\sigma_{12}^2)=\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{z}_i^2}{\sigma_i^2}\cdot\frac{1}{2+\frac{\sigma_i^2}{\sigma_{12}^2}+\frac{\sigma_{12}^2}{\sigma_i^2}}-1=0. \quad (27)$$

Now we proceed to prove the conclusions in Theorem 1.

Denote the left-hand side of Eq. 15 in Theorem 1 as d and plug Eq. 26 in, thus Eq. 15 means

$$d=\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{z}_i^2}{\sigma_i^2}>2+\sqrt{\frac{\sigma_1^2}{\sigma_n^2}}+\sqrt{\frac{\sigma_n^2}{\sigma_1^2}}\geq 4.$$

When $\sigma_{12}^2=k_1=(d-2)\sigma_1^2$, for any i, we have

$$\frac{1}{2+\frac{\sigma_i^2}{\sigma_{12}^2}+\frac{\sigma_{12}^2}{\sigma_i^2}}<\frac{1}{2+0+d-2}=\frac{1}{d}.$$

Thus,

$$F(k_1)<\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{z}_i^2}{\sigma_i^2}\cdot\frac{1}{d}-1=0.$$

When $\sigma_{12}^2=k_2=\sqrt{\sigma_1^2\sigma_n^2}$, for any i,

$$\frac{1}{2+\frac{\sigma_i^2}{\sigma_{12}^2}+\frac{\sigma_{12}^2}{\sigma_i^2}}\geq\frac{1}{2+\frac{\sigma_n^2}{k_2}+\frac{k_2}{\sigma_1^2}}=\frac{1}{2+\sqrt{\frac{\sigma_1^2}{\sigma_n^2}}+\sqrt{\frac{\sigma_n^2}{\sigma_1^2}}}\geq\frac{1}{d},$$

thus

$$F(k_2)\geq\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{z}_i^2}{\sigma_i^2}\cdot\frac{1}{d}-1=0.$$

Since $0<k_2<k_1$ and $F(\cdot)$ is continuous, there must exist a k_3 such that $k_2\leq k_3<k_1$ and $F(k_3)=0$. This proves that the inequality Eq. 15 in Theorem 1 holds can indicate a non-zero solution for Eq. 24, namely there exists at least one non-zero fixed point for σ_{12}^2 in the Bayesian EM, which means the two Gaussian sources are not consensus according to our definition of inconsistency. Thus, the first claim in Theorem 1 is proved.

Eq. 16 means $d=\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{z}_i^2}{\sigma_i^2}<4$, then we have

$$F(\sigma_{12}^2)\leq\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{z}_i^2}{\sigma_i^2}\cdot\frac{1}{4}-1=\frac{d}{4}-1<0,$$

for all $\sigma_{12}^2>0$. Therefore, there does not exist a non-zero solution for Eq. 27. Eq. 16 in Theorem 1 is proven. □

ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation Grants IIS-0347877 and IIS-0308222, and OMRON Corp.

REFERENCES

[1] M. Isard and A. Blake, "Contour tracking by stochastic propagation of conditional density," in *European Conf. on Computer Vision (ECCV'96)*, Cambridge, UK, Apr. 1996, pp. 343–356.

[2] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'98)*, Santa Barbara, CA, June 23-25, 1998, pp. 232 – 237.

[3] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 5, pp. 564 – 577, May 2003.

[4] M. J. Black and A. D. Jepson, "Eigentracking: Robust matching and tracking of articulated objects using a view-based representation," in *European Conf. on Computer Vision (ECCV'96)*, Cambridge, UK, Apr. 1996, pp. 329–342.

[5] G. Hager and P. Belhumeur, "Real-time tracking of image regions with changes in geoemtry and illumination," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'96)*, San Francisco, CA, June 18-20, 1996, pp. 403–410.

[6] G. Hager, M. Dewan, and C. Stewart, "Multiple kernel tracking with SSD," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'04)*, vol. 1, Washington, DC, June 27 - July 2, 2004, pp. 790–797.

[7] K.-C. Lee and D. Kriegman, "Online learning of probabilistic appearance manifolds for video-based recognition and tracking," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, San Diego, CA, June 20-25, 2005, pp. 852 – 859.

[8] J. Lim, D. Ross, R.-S. Lin, and M.-H. Yang, "Incremental learning for visual tracking," in *Advances in Neural Information Processing Systems 17 (NIPS'04)*, Vancouver, Canada, Dec. 13-18, 2004, pp. 801–808.

[9] M. Yang and Y. Wu, "Tracking non-stationary appearances and dynamic feature selection," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, San Diego, CA, June 20-25, 2005, pp. 1059 – 1066.

[10] S. Avidan, "Support vector tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, pp. 1064 – 1072, Aug. 2004.

[11] ——, "Ensemble tracking," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, San Diego, CA, June 20-25, 2005, pp. 494 – 501.

[12] J. Wang, X. Chen, and W. Gao, "Online selecting discriminative tracking features using particle filter," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, San Diego, CA, June 20-25, 2005, pp. 1037 – 1042.

[13] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94)*, Santiago, Chile, 12-15, 1994, pp. 487 – 499.

[14] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *IEEE Int'l Conf. on Computer Vision (ICCV'03)*, vol. 2, Nice, France, Oct. 13-16, 2003, pp. 1470 – 1477.

[15] ——, "Video data mining using configurations of viewpoint invariant regions," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'04)*, vol. 1, Washington, DC, June 27 - July 2, 2004, pp. 488 – 495.

[16] M. Leordeanu and R. Collins, "Unsupervised learning of object features from video sequences," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, San Diego, CA, June 20-25, 2005, pp. 1142 – 1149.

[17] X. S. Zhou, D. Comaniciu, and A. Gupta, "An information fusion framework for robust shape tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 1, pp. 115–129, Jan. 2005.

[18] D. G. C. Race, "http://www.darpa.mil/grandchallenge."

[19] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185 – 203, 1981.

[20] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *DARPA Image Understanding Workshop*, Apr. 1981, pp. 121 – 130.

[21] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *European Conf. on Computer Vision (ECCV'04)*, vol. 4, Prague, Czech Republic, May 2004, pp. 25 – 36.

[22] S. Roth and M. J. Black, "On the spatial statistics of optical flow," in *IEEE Int'l Conf. on Computer Vision (ICCV'05)*, vol. 1, Beijing, China, Oct. 2005, pp. 42 – 49.

[23] A. Jepson, D. Fleet, and T. El-Maraghi, "Robust online appearance models for visual tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 10, pp. 1296 – 1311, Oct. 2003.

[24] D. Ramanan, D. A. Forsyth, and A. Zisserman, "Tracking people by learning their appearance," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, no. 1, pp. 65 – 81, Jan. 2007.

[25] K. Okuma, A. Taleghani, N. D. Freitas, J. J. Little, and D. G. Lowe, "A boosted particle filter: Multitarget detection and tracking," in *European Conf. on Computer Vision (ECCV'04)*, vol. 1, Prague, Czech Republic, May 2004, pp. 28 – 39.

[26] T. Zhao and R. Nevatia, "Tracking multiple humans in complex situations," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 9, pp. 1208 – 1221, Sept. 2004.

[27] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'99)*, vol. 2, Fort Collins, CO, June 23-25, 1999, pp. 246 – 252.

[28] J. Shi and J. Malik, "Motion segmentation and tracking using normalized cuts," in *IEEE Int'l Conf. on Computer Vision (ICCV'98)*, Bombay, India, Jan. 1998, pp. 1154 – 1160.

[29] O. Williams, A. Blake, and R. Cipolla, "Sparse bayesian learning for efficient visual tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 8, pp. 1292 – 1304, Aug. 2005.

[30] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, no. 4, p. 13, Dec. 2006.

[31] M. Isard and A. Blake, "CONDENSATION - conditional density propagation for visual tracking," *Int'l Journal of Computer Vision (IJCV)*, vol. 29, pp. 5 – 28, May 1998.

[32] D. B. Reid, "An algorithm for tracking multiple targets," *IEEE Trans. Automat. Contr.*, vol. 24, no. 6, pp. 843 – 854, Dec. 1979.

[33] I. J. Cox and S. L. Hingorani, "An efficient implementation of reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, no. 2, pp. 138–150, Feb. 1996.

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

[35] C. Rasmussen and G. D. Hager, "Probabilistic data association methods for tracking complex visual objects," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 6, pp. 560–576, June 2001.

[36] Y. Wu and T. S. Huang, "Robust visual tracking by integrating multiple cues based on co-inference learning," *Int'l Journal of Computer Vision (IJCV)*, vol. 58, no. 1, pp. 55–71, June 2004.

[37] R. T. Collins, Y. Liu, and M. Leordeanu, "Robust online appearance models for visual tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 10, pp. 1631 – 1643, Oct. 2005.

[38] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via online boosting," in *The British Machine Vision Conference (BMVC'06)*, vol. 1, Edinburgh, 4-7, 2006, pp. 47 – 56.

[39] M. Yang, Y. Wu, and S. Lao, "Intelligent collaborative tracking by mining auxiliary objects," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, NYC, June 17-22, 2006, pp. 697 – 704.

[40] J. Shi and C. Tomasi, "Good features to track," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'94)*, Seattle, WA, June21 - 23, 1994, pp. 593 – 600.

[41] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE Int'l Conf. on Computer Vision (ICCV'99)*, vol. 2, Corfu, Greece, Sept. 20-27, 1999, pp. 1150 – 1157.

[42] K. Mikolajczyk and C. Schmid, "Indexing based on scale invariant interest points," in *IEEE Int'l Conf. on Computer Vision (ICCV'01)*, vol. 1, Vancouver, Canada, July 7 - 14, 2001, pp. 525 – 531.

[43] A. Fitzgibbon and A. Zisserman, "On affine invariant clustering and automatic cast listing in movies," in *European Conf. on Computer Vision (ECCV'02)*, vol. 3, Copenhagen, Denmark, May 27 - June 2, 2002, pp. 304 – 320.

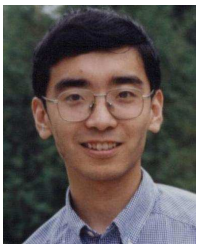
[44] R. Jain, R. Kasturi, and B. G. Schunck, *Machine Vision*. McGrawHill, Inc, 1995.

[45] S. J. Julier and J. K. Uhlmann, "A non-divergent estimation algorithm in the presence of unknown correlations," in *Proceedings of the American Control Conference (ACC'97)*, Albuquerque, New Mexico, June 4-6, 1997, pp. 2369 – 2373.

[46] G. Hua and Y. Wu, "Measurement integration under inconsistency for robust tracking," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, NYC, June 17-22, 2006, pp. 650– 657.

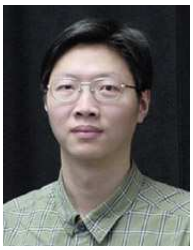
[47] EC funnded CAVIAR project/IST 2001 37540, "http://homepages.inf.ed.ac.uk/rbf/caviar/."

[48] V. I. Pavlovic, "Dynamic bayesian networks for information fusion with application to human-computer interfaces," Ph.D. dissertation, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, IL, 1999.



Ming Yang received the B.E. and M.E. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2001 and 2004, respectively, and the PhD degree in electrical and computer engineering from Northwestern University, Evanston, Illinois, in June 2008.

From 2004 to 2008, he was a research assistant of Prof. Ying Wu in the Computer Vision group of Northwestern University. He will join NEC Laboratory, America, Cupertino, California, as a member of research staff upon his graduation. His research interests include computer vision, machine learning, video communication, medical image analysis, and intelligent multimedia content analysis. He was an excellent bachelor graduate of Tsinghua University, 2001. He was also awarded the excellent student fellowship from 1998 to 2003 at Tsinghua University. He is a member of the IEEE.



Ying Wu received the B.S. from Huazhong University of Science and Technology, Wuhan, China, in 1994, the M.S. from Tsinghua University, Beijing, China, in 1997, and the Ph.D. in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Urbana, Illinois, in 2001.

From 1997 to 2001, he was a research assistant at the Beckman Institute for Advanced Science and Technology at UIUC. During summer 1999 and 2000, he was a research intern with Microsoft Research, Redmond, Washington. In 2001, he joined the Department of Electrical and Computer Engineering at Northwestern University, Evanston, Illinois, as an assistant professor. He is currently an associate professor of Electrical Engineering and Computer Science at Northwestern University. His current research interests include computer vision, image and video analysis, pattern recognition, machine learning, multimedia data mining, and human-computer interaction. He serves as associate editors for IEEE Transactions on Image Processing, SPIE Journal of Electronic Imaging, and IAPR Journal of Machine Vision and Applications. He received the Robert T. Chien Award at UIUC in 2001, and the NSF CAREER award in 2003. He is a senior member of the IEEE.

Gang Hua is a Scientist at Microsoft Live labs Research. He obtained his Ph.D. degree in Electrical and Computer Engineering from Northwestern University in June 2006. His current research interests include computer vision, machine learning, visual recognition, intelligent image/video/multimedia processing, visual motion and content analysis, and their applications to multimedia search.

He was a research assistant of Prof. Ying Wu in the Computer Vision group of Northwestern University from 2002 to 2006. During the summer 2005 and summer 2004, he was a research intern with the Speech Technology Group, Microsoft Research, Redmond, WA, and a research intern with the Honda Research Institute, Mountain View, CA, respectively. Before coming to Northwestern, he was a research assistant in the Institute of Artificial Intelligence and Robotics at Xian Jiaotong University (XJTU), Xian, China. He received his M.S. in pattern recognition and intelligence system at XJTU in 2002. He was enrolled in the Special Class for the Gifted Young of XJTU in 1994 and received his B.S. in Automatic Control Engineering in 1999. He received the Richter Fellowship and the Walter P. Murphy Fellowship at Northwestern University in 2005 and 2002, respectively. When he was in XJTU, he was awarded the Guanghua Fellowship, the Eastcom Fellowship, the Most Outstanding Student Exemplar Fellowship, the Sea-star Fellowship and the Jiangyue Fellowship in 2001, 2000, 1997, 1997 and 1995 respectively. He was also a recipient of the University Fellowship from 1994 to 2001 at XJTU. He is a member of IEEE. As of May, 2008, he holds 1 US patent and has 10 more patents pending.