Decentralized Multiple Target Tracking using Netted Collaborative Autonomous Trackers

Ting Yu and Ying Wu
Department of Electrical & Computer Engineering
Northwestern University
2145 Sheridan Road, Evanston, IL 60208
{tingyu,yingwu}@ece.northwestern.edu

Abstract

This paper presents a decentralized approach to multiple target tracking. The novelty of this approach lies in the use of a set of autonomous while collaborative trackers to overcome the tracker coalescence problem with linear complexity. In this approach, the individual trackers are autonomous in the sense that they can select targets to track and evaluate themselves, and they are also collaborative since they need to compete for the targets against those trackers that are close to them through communication. The theoretical foundation of this new approach is based on the variational analysis of a Markov network that reveals the collaborative mechanism through a fixed point iteration among these trackers and the existence of the equilibriums. In addition, a trained object detector is incorporated to help sense the potential newly appearing targets in the dynamic scene. Experimental results on challenging video sequences demonstrate the effectiveness and efficiency of the proposed method.

1 Introduction

A major difficulty of multiple target tracking lies in the fact that the tracker are insensitive to the differences among the targets such that they may not be distinguishable from each other, which leads to a combinatorial problem on target-tracker association. We can call it as the "identical" targets problem. The neglect of this problem (e.g., by using multiple independent tracker) will generally lead to the tracker coalescence phenomenon, i.e., several trackers are associated to one same target while other targets lose track. Coalescence often takes place especially when the targets are close or present occlusions [9, 20].

Most existing solutions to this problem are based on the centralized methodology by considering joint data association. Due to the exploring of a high-dimensional joint state space, these methods are generally computational intensive. For example, the multiple hypothesis tracker (MHT) [2, 5] and the joint probabilistic data association filter (JPDAF) [13] have to exhaust all possible associations, and the sampling-based methods [6, 7, 11, 16, 21] demand a huge number of particles.

Such centralized solutions are fine to a powerful processor. However, they are not appropriate for the emerging application of sensor networks, in which there are a large number of sensing units that have the functionality of sensing, computing and communicating. However, these units are power-limited to prevent much computation and communication [10]. Thus, to make good use of such sensor networks for target tracking, complex computation must be distributed into the network, since once a certain unit takes charge of sensing, its computational load on target tracking needs to be migrated to other idle units. Although this research is being carried out at the computer architecture level, it is more desirable to find a decentralized scheme at the algorithm-level for efficient tracking of multiple targets, since it will leads to the essential parallelization and distributed computing.

In this paper, we present a new and efficient decentralized visual tracking method for multiple "identical" targets. The novelty of our approach lies in the use of a set of autonomous while collaborative trackers to overcome the tracker coalescence problem with linear complexity. In this approach, the individual trackers are autonomous in the sense that they can select targets to track and evaluate themselves, and they are also collaborative since they need to compete for the targets against those trackers that are close to them through communication.

The theoretical foundation of the new approach is based on a Markov network formulation, where each hidden node in the network represents the state of an autonomous tracker. The trackers can be either in active or inactive status, indicating if they are currently following targets or not. The trackers are self-aware since they can determine their own status by an entropy-based evaluator. The edges of the network represent the constraints among these individual trackers induced by the target competition. The structure of the network keeps changing with the states of the track-

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 **IEEE**

ers. The collaboration mechanism of these netted trackers is revealed by the information exchanges of these trackers in a fixed point-like iteration that reaches equilibriums, based on the variational analysis of the Markov network.

In addition, a roughly trained AdaBoost-based target detector [18] is equipped to each tracker to help sense the potential newly appearing targets in the dynamic scene, therefore background subtraction is not necessary to our method, although it can largely help the case of fixed backgrounds. The use of object detectors within each tracker also supports the construction of an effective importance function, which leads to a more effective variational inference. Extensive experiments on the challenging video sequences are conducted to demonstrate the effectiveness and efficiency of the proposed method.

2 Related Work

Many multiple target tracking methods have been developed during the past few years, where most of them are based on the centralized joint state space inference either under the parametric or non-parametric formulations. The parametric methods, such as multiple hypothesis tracking (MHT) [2, 5] and joint probabilistic data association filtering (JPDAF) [13] handle the coalescence problem by the joint data association principle in which one image observation can only support a single target hypothesis and one target hypothesis can only occupy a single observation, therefore suffering from the combinatorial complexity due to the exhaustive enumeration for all possible associations. Based on Monte Carlo sampling techniques, non-parametric methods [6, 7, 11, 16, 21] can tackle the coalescence problem in a top-down process that generates and evaluates a large number of hypotheses, thus also confronted by a similar high computational cost due to the exponential demand of the increase of particles. All these approaches are actually dealing with the centralized state space directly, which results in the inevitable combinatorial or exponential complexity in the algorithm level that is hardly scalable.

The existing approaches can also be classified into two categories according to whether a fixed background model is employed. Background subtraction normally offers a strong localization clue for detecting each new target entering the scene. Whenever a new target is appearing, a new tracker can be immediately instantiated to follow it [4]. The fixed background assumption is also the essential reason why the configuration level optimization techniques, such as jump-diffusion Markov chain Monte Carlo in [21] and variants of particle filtering [7, 16], can be applied to inference the number of targets existing in the scene over the union of joint state space of multiple targets, since under this assumption the observation likelihood can be calculated based on the whole image information. Foreground area is evaluated by the foreground target model, background area is also assessed by the maintained background model, which in combination makes the configuration level reasoning feasible. However, this nice property does not exist under the changing background situations, since there is generally no way to maintain a powerful background model to explain all non-target areas in the dynamic scene. Therefore existing approaches dealing with the dynamic background scenarios are either assuming to track fixed number of targets [6, 11, 13, 20], which obviously limits its generalization, or adopting an target detector to help determine if any new target appears in the scene [12]. However, in [12] it stays unclear how the coalescence problem can be reliably solved under their single particle filtering framework.

Different from these existing methods, we proposes a decentralized approach to multiple target tracking by using a set of collaborative autonomous trackers. Compared to the state of the art, this new approach proves to be computationally efficient and algorithm-level parallelizable.

3 Collaborative Trackers

As indicated in Sec. 1, the methods based on a set of independent trackers (denoted as M.i.T. methods) are insufficient to the task of multiple target tracking, especially when the targets are similar. This difficulty is not uncommon since the image observation model used in the trackers may not distinguish the nuance within a class of objects. Thus, the tracker *coalescence* problem occurs, when several targets are close or when they occlude each other. As indicated in [20], the root of *coalescence* lies in the violation of the independence assumption due to the so-called conditional dependency induced by the mixed image observations.

Different from the centralized methodology that considers the joint data associations, the proposed approach consists of a set of collaborative autonomous trackers, each of which copes with a single target. These trackers are autonomous since they, by themselves, search for targets to track and evaluate themselves (Sec. 3.2). At the same time, they are collaborative since they communicate, exchange intelligence and cooperate (Sec. 3.3). These two mechanisms are integrated under the formulation of a Markov Network (Sec. 3.1).

3.1 A Tracker Network Formulation

We denote the state (i.e., the motion to be estimated) of each individual tracker at time t by $\mathbf{x}_{i,t}$, its associated image observation by $\mathbf{z}_{i,t}$, the joint target states by $\mathbf{X_t} = \{\mathbf{x}_{1,t}, \ldots, \mathbf{x}_{M,t}\}$, and the joint observation by $\mathbf{Z_t}$ for a set of M trackers. We denote $\underline{\mathbf{Z}}_t = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_t\}$.

The problem of estimating the posterior of the joint state $\mathbf{X}_t$ from image measurements $\underline{\mathbf{Z}}_t$ can be casted as a Bayesian inference problem of a Markov Network, as illustrated in Figure 1, where the circles are the hidden variables and the squares are the evidence variables.

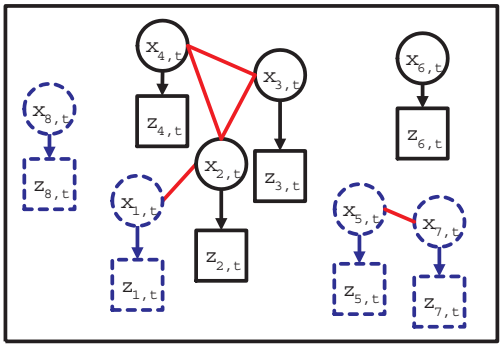

Figure 1: Collaborative trackers as a Markov network.

Each pair of a circle node (i.e., $\mathbf{x}_{i,t}$) and a square node (i.e., $\mathbf{z}_{i,t}$) represents an individual tracker, in which the directed link connecting them represents the image observation model $p(\mathbf{z}_{i,t}|\mathbf{x}_{i,t})$. These trackers can switch between two modes: *active* (shown as solid) or *inactive* (shown as dotted). When a tracker is following a target, it is active; and inactive otherwise. The details of tracker self-evaluation is given in Sec. 3.2. Thus, we write the joint conditional likelihood as:

$$p(\mathbf{Z}_t|\mathbf{X}_t) = \prod_{i=1}^{M} p_i(\mathbf{z}_{i,t}|\mathbf{x}_{i,t}) \quad (1)$$

At the same time, these individual trackers are correlated (shown as the links in the graph) when their state variables are constrained. For example, when several trackers are close, they are competing targets for tracking, thus enforcing an exclusive constraint that a target must not be associated to more than one tracker. If trackers are far apart enough, such a constraint does not apply and the problem degenerates to the independent case. Therefore, the structure of the Markov network changes during tracking. It is worth mentioning that the pair-wise state constraint model has also been successfully applied in articulated hand tracking based on a similar markov network formulation [15].

The constraints apply to the individual trackers no matter whether they are currently active or inactive. We model such constraints as a motion prior term $p(\mathbf{X}_t)$, and represent it as a general Gibbs distribution:

$$p(\mathbf{X}_t) = \frac{1}{Z_c} \prod_{i \in V} \psi_i(\mathbf{x}_{i,t}) \prod_{(i,j) \in E} \psi_{ij}(\mathbf{x}_{i,t}, \mathbf{x}_{j,t}) \quad (2)$$

where V denotes the set of nodes and E the set of edges in the graph, $\psi_i(\mathbf{x}_{i,t})$ is the local prior for tracker $\mathbf{x}_{i,t}$ and is explicitly modelled as the dynamic prior propagated from previous time instance, i.e., $\psi_i(\mathbf{x}_{i,t}) \propto p(\mathbf{x}_{i,t}|\underline{\mathbf{Z}}_{t-1})$, $\psi_{ij}(\mathbf{x}_{i,t}, \mathbf{x}_{j,t})$ is a potential function stipulating the motion constraints between neighboring nodes $\mathbf{x}_{i,t}$ and $\mathbf{x}_{j,t}$. To model the above competition (or exclusive) constraints, the potential function shall have a smaller value when the pair of trackers become closer, such that it is less likely of having crowded trackers. As a special design, this potential function can be modelled as follows:

$$\psi_{ij}(\mathbf{x}_{i,t}, \mathbf{x}_{j,t}) \propto e^{d(\mathbf{x}_{i,t}, \mathbf{x}_{j,t})^T \Sigma_{ij}^{-1} d(\mathbf{x}_{i,t}, \mathbf{x}_{j,t})} \quad (3)$$

where $d(\mathbf{x}_{i,t}, \mathbf{x}_{j,t}) = \mathbf{x}_{i,t} - \mathbf{x}_{j,t}$ is the difference of the state variables of the competitive trackers, and Σ_{ij} characterizes the size of possible competition region in the sate space.

By this means, the coalescence problem can be largely prevented with the introduction of the competition mechanism among the trackers, and the capturing of the newly-appeared targets can be fulfilled as well. For example, when the competition presents among active trackers, such a potential term acts to overcome the coalescence problem as described before. When the competition presents among inactive trackers, this potential term helps to force these inactive trackers to search different image regions for newly appearing targets to track. When the competition happens between an active tracker and an inactive one, such an elastically exclusive force becomes unidirectional, i.e. only the active tracker can exclude the inactive one to prevent the case where the inactive tracker "hijacks" the target being tracked by the active one. These mechanisms are explicitly formulated as the priors in Eq. 2, and play an important role in the collaborations among the set of individual trackers.

3.2 Self-awareness and Mode Switching

At each time instant t, the structure of the tracker network is determined according to the relative positions of the trackers, calculated by the conditional mean state estimator $\overline{\mathbf{x}}_t = \int \mathbf{x}_t p(\mathbf{x}_t|\underline{\mathbf{z}}_t) d\mathbf{x}_t$.

Many existing multiple target tracking approaches assume fixed backgrounds [4, 5, 7, 16, 21], since the pixel level likelihood facilitates efficient detection of the appearing and disappearing of the targets. In this paper, we do not limit our approach to this assumption. Therefore, to make possible the capturing of the new targets in dynamic video scenes, each autonomous tracker is equipped with a rough local range detector that only searches its nearby regions. When a new target enters the video scene, it may not be immediately sensed and tracked by any of the trackers due to their limited monitoring areas. But their collaboration will gradually distribute them to cover the entire image region such that the new targets can be eventually detected and tracked. In general, the process of pickup is quick in several frames, depending on the size of the tracker network. This is also validated in our experiments. Although this may induce detection lag, it saves computation significantly. An extreme case is to set the detection range of each tracker to be the entire image to obtain instant detection, but incurring demanding computation. Thus, in practice, we need to balance between the detection lag and the computational cost.

As described in Sec. 3.1, individual trackers are autonomous and should be able to evaluate themselves. They need to determine and switch the modes (i.e., active or inactive) by themselves. We denote by $\mathbf{r}_t$ the binary performance indicator for each tracker.

Based on the inference result $p(\mathbf{x}_t|\underline{\mathbf{z}}_t)$, there may exist different ways to obtain a performance indicator for each tracker. We observe that when each single tracker is experiencing good tracking conditions, the underlying posterior $p(\mathbf{x}_t|\underline{\mathbf{z}}_t)$ will mainly demonstrate some sharp unimode distribution. On the contrary, a more uniform prior implies larger uncertainty of the motion estimation, i.e., the tracking result is less confident and thus not satisfactory. Therefore, an entropy measure can be used as a good performance metric to evaluate the tracking performance. Specifically, we define the performance indicator $\mathbf{r}_t$ as follows:

$$\mathbf{r}_t = \begin{cases} 1 & \text{if } -\int p(\mathbf{x}_t|\underline{\mathbf{z}}_t)\mathbf{log}p(\mathbf{x}_t|\underline{\mathbf{z}}_t) < \tau \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\mathbf{r}_t = 1$ indicates active trackers since the target seems to be successfully followed by the tracker, while $\mathbf{r}_t = 0$ implies inactive trackers otherwise, such as the tracker loses track due to the interferences from the cluttered background or simply because the previously tracked target leaves the video scene. The threshold τ can be empirically determined.

The modes of an autonomous tracker determine its dynamics model $p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1})$ (where i indexes the tracker). Thus a track can switch its behaviors autonomously based on the active or inactive mode determined by itself:

$$p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1}) = \begin{cases} p_a(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1}) & \text{if } \mathbf{r}_{i,t-1} = 1 \\ p_u(\mathbf{x}_{i,t}|\overline{\mathbf{x}}_{i,t-1}) & \text{otherwise} \end{cases} \tag{5}$$

where $p_a(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1})$ is a constant acceleration motion model, and $p_u(\mathbf{x}_{i,t}|\overline{\mathbf{x}}_{i,t-1})$ is a uninformative uniformly random walk around the tracker's previous conditional mean state estimator $\overline{\mathbf{x}}_{i,t-1}$.

Camouflages may affect the tracking performance significantly, no matter whether these camouflages arise from the same types of targets in the nearby region or simply due to the background clutters that resemble the target. Thus, it is not robust when tracking one target. This difficulty can be largely alleviated by the joint tracking of multiple similar targets since the joint data association can largely reduce the risk of loss track of any. Casting this idea into a decentralized methodology, we believe the collaborations among individual trackers act as a distributed way for data association.

3.3 Collaboration and Decentralization

How can the trackers in the network collaborate? Is there an optimal collaboration strategy? We present in this section a theoretical foundation that supports our proposed collaborative tracker.

3.3.1 Variational Analysis and Decentrailization

With the modelling of $p(\mathbf{Z}_t|\mathbf{X}_t)$ and $p(\mathbf{X}_t)$ in Sec. 3.1, the joint posterior of the tracker is given by:

$$\begin{aligned} p(\mathbf{X}_t|\underline{\mathbf{Z}}_t) &\propto p(\mathbf{Z}_t|\mathbf{X}_t)p(\mathbf{X}_t) \\ &\propto \prod_{i\in V} p_i(\mathbf{z}_{i,t}|\mathbf{x}_{i,t})p(\mathbf{x}_{i,t}|\underline{\mathbf{Z}}_{t-1}) \prod_{(i,j)\in E} \psi_{ij}(\mathbf{x}_{i,t},\mathbf{x}_{j,t}) \end{aligned} \tag{6}$$

It seems infeasible to calculate such a complicated posterior in a direct manner, since it involves multiple dimensional integration. Because it is very likely that the Markov networks in our formulation contain loops when three or more trackers are linked together, belief propagation [3] may also not be appropriate here. In contrast to belief propagation, probabilistic variational analysis [8, 19, 20] can be employed for approximation, especially for loopy networks.

As in [20], a fully factorized variational density $Q(\mathbf{X}_t) = \prod_{i\in V} Q_i(\mathbf{x}_i)$ can be used to approximate the true posterior $p(\mathbf{X}_t|\underline{\mathbf{Z}}_t)$. We can find the optimal approximation in the sense that the Kullback-Leibler (KL) divergence between this variational distribution and the posterior is minimized, i.e.:

$$Q^*(\mathbf{X}_t) = \arg\min_Q KL(Q(\mathbf{X}_t)||p(\mathbf{X}_t|\underline{\mathbf{Z}}_t)) \tag{7}$$

The solution of this optimization problem is given by the following fixed point equation [19]:

$$\begin{aligned} Q_{i,t}(\mathbf{x}_{i,t}) \quad \longleftarrow \quad & \frac{1}{Z_i'} p_i(\mathbf{z}_{i,t}|\mathbf{x}_{i,t}) \\ \times \quad & \int p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1})p(\mathbf{x}_{i,t-1}|\underline{\mathbf{z}}_{i,t-1}) \\ \times \quad & M_{i,t}(\mathbf{x}_{i,t}) \end{aligned} \tag{8}$$

where $M_{i,t}(\mathbf{x}_{i,t})$ is derived as

$$M_{i,t}(\mathbf{x}_{i,t}) = \exp\{\sum_{j\in\mathcal{N}(i)} \int_{\mathbf{x}_{j,t}} Q_{j,t}(\mathbf{x}_{j,t}) \log \psi_{ij}(\mathbf{x}_{i,t},\mathbf{x}_{j,t})\}, \tag{9}$$

where Z_i' is a constant partition function, and $\mathcal{N}(i)$ is the neighborhood of the tracker i.

Eq. 8 shows that posterior of a tracker i is determined by three factors: its own image likelihood $p_i(\mathbf{z}_{i,t}|\mathbf{x}_{i,t})$ measure, its own dynamics predictions $p(\mathbf{x}_{i,t}|\underline{\mathbf{z}}_{i,t-1})$, and more importantly, the "collaborative message" $M_{i,t}(\mathbf{x}_{i,t})$ passed from its neighborhood trackers $x_{\mathcal{N}(i),t}$ that compete the common image resources against it. It can be shown that the computational complexity of the collaborative tracking based on this fixed point iterations is linear with respect to the number of netted trackers, which is actually a significant improvement over the methods that work directly on the joint state spaces. In Eq. 8, it is clear that the basic computation unit is the posterior estimation for each individual

tracker, therefore, the computationally demanding tracking task in Eq. 6 has been decentralized to the set of netted autonomous trackers with the cost of communication and collaboration.

3.3.2 Mixture Density of Importance Sampling

Considering the fact that the posteriors of each tracker may not be Gaussian, the above fixed point collaboration can be well implemented based on sequential Monte Carlo technique, in which a particle set is employed to represent the variational density $Q_{i,t}(\mathbf{x}_{i,t})$, i.e.,

$$Q^k_{i,t}(\mathbf{x}_{i,t}) \sim \{s^{(n)}_{i,t}(k), \pi^{(n)}_{i,t}(k)\}^N_{n=1} \tag{10}$$

where s and π denote the sample and its weight, N is the number of samples, and k is the iteration index during the fixed point iterations.

Although we can generate the new sample set $\{s^{(n)}_{i,t}, \pi^{(n)}_{i,t}\}^N_{n=1}$ of tracker i for the current time instant t solely from its dynamics model $p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1})$, a better way to achieve the effective sampling is to incorporate the bottom-up image information at the current frame into consideration, and design informative importance functions to guide the particle sampling [12, 14]. Since in our formulation, each autonomous tracker is equipped with a local region object detector, which obviously facilitates the tracker to collect the bottom up information to construct an effective importance sampling. At time t, tracker i detects $\mathcal{C}$ potential targets within its monitoring area, and each of these detections is depicted by the detected location and scale $\{\mathbf{O}_{c,t}, c \in \mathcal{C}\}$. We construct the following mixture density as an effective importance function:

$$\begin{aligned} I_{i,t}(\mathbf{x}_{i,t}) \quad &= \alpha[\textstyle\sum_{c=1}^{\mathcal{C}} \omega_{c,t} N(\mathbf{x}_{i,t}|\mathbf{O}_{c,t}, \sum_{c,t})] \\ &\quad +(1-\alpha)[p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1})] \end{aligned} \tag{11}$$

where N is a Gaussian density with mean vector $\mathbf{O}_{c,t}$ and diagonal covariance matrix $\sum_{c,t}$, the Gaussian mixture weights $\omega_{c,t}$ are empirically determined based on the detection's location relative to the current position of the tracker. The parameter α balances between target detection and tracker's dynamics. When the tracker is under good conditions, α should be small and the sensing region for tracker's target detection will also be reduced such that the dynamics model plays the dominant role. On the other hand, if the tracker is experiencing a tracking failure or unable to detect anything, α will become large and the detection region will also expand to facilitate the search of the lost target or any potential new targets.

Therefore, based on Eq. 8, the sequential Monte Carlo implementation of the proposed collaborative tracker can be summarized as in Figure 2.

1. Structure Determination of Markov Network:
   At time t, determine the Markov network structure according to the relative positions and performance indicators of the trackers from time $t-1$.
2. Importance Sampling:
   For tracker $i, i \in M$, generate new samples $\{s^{(n)}_{i,t}\}^N_{n=1}$ from importance function $I_{i,t}(\mathbf{x}_{i,t})$.
3. Importance Re-weighting:
   For each $s^{(n)}_{i,t}$, set its re-weight
   $\widetilde{\omega}^{(n)}_{i,t} = [\sum_{m=1}^N \pi^{(m)}_{i,t-1} p(s^{(n)}_{i,t}|s^{(m)}_{i,t-1})]/I_{i,t}(s^{(n)}_{i,t})$
4. Observation Likelihood Calculation:
   For each $s^{(n)}_{i,t}$, perform likelihood calculation
   $w^{(n)}_{i,t} = p(z_{i,t}|s^{(n)}_{i,t})$
5. Iteration: Initially set $k=0$ and $k = k+1$;
   (a) calculate the "message" from neighbors:
   $$m^{(n)}_{i,t}(k) = \sum_{j \in \mathcal{N}(i)} \sum_{m=1}^{N} \pi^{(m)}_{j,t}(k-1) \log \psi_{ij}(s^{(n)}_{i,t}, s^{(m)}_{j,t}).$$
   (b) Re-weight the particles by:
   $$\pi^{(n)}_{i,t}(k) = e^{m^{(n)}_{i,t}(k)} \cdot w^{(n)}_{i,t} \cdot \widetilde{\omega}^{(n)}_{i,t}.$$
   (c) normalize to obtain
   $$Q^k_{i,t}(\mathbf{x}_{i,t}) \sim \{s^{(n)}_{i,t}, \pi^{(n)}_{i,t}(k)\}$$

Figure 2: The Monte Carlo implementation of the proposed collaborative tracker.

4 Experiments

The proposed approach is implemented to perform experiments on tracking sports players in real-life video sequences of soccer and hockey games. In both these experiments, a set of 16 trackers is casted to cover the changing background scenes. Each tracker is equipped with an object detector, which is trained using AdaBoost, to help sense the potential appearing sports players within its local range. The training data of the detector for each testing sequence is collected by manually labelling the sports players regions from randomly selected 50 frames of that sequence.

The individual tracker is a rectangle region tracker, where the target state $\mathbf{x}_i$ is modelled by 2D similarity transformation parameters, i.e. translation and scale. The dynamics model $p(\mathbf{x}_{i,t}|\mathbf{x}_{i,t-1})$ is either a constant acceleration model or a uniformly random walk model depending on the performance indicator $\mathbf{r}_{i,t-1}$, as described in section 3.2. The likelihood function $p(\mathbf{z}_i|\mathbf{x}_i)$ is a color-histogram based observation model built in HSV color space, which is known insensitive to illumination changes. The histogram model of the target is also trained using the same data set as of training AdaBoost detector. 100 particles are used to represent the posterior of each tracker, which leads to only 1600 particles in total, linear with the number of trackers, to monitor the appearing and disappearing of sports players in the highly dynamic scenes. Under this parameter setting, our collaborative trackers runs

around 15 fps on a $P4$ 2GHz PC (Note that our decentralized scheme is theoretically parallel at the algorithm-level, therefore with code optimization, much faster performance can be easily achieved).

We firstly test the proposed approach on tracking multiple soccer players in a video sequence of soccer match, in which the appearing and disappearing of players often happen along the sequence. The maximum number of players simultaneously presenting in the scene is 8. Note that in this sequence there are two team players, one is wearing white sports clothes, while the other is wearing red. Therefore, our 16 trackers are equally divided into two sets, which share the same trained object detector but have different image likelihood functions, each specifically trained for one team.

The gradually detecting of the present soccer players in the viewing scene is shown in Figure 3 that are corresponding to the 2, 18, 29 frames of the sequence respectively. The casted autonomous trackers, which are inactive initially, start to roam around the scene with random walk to sense their potential targets, while at the same time forcing their neighboring inactive ones to search around other unchecked areas by communicating with them through variational message, as shown in Figure 2. In general it will lead to a roughly uniform coverage over the whole image area as can be seen in Figure 3. The red thick rectangles in the Figure illustrate the active trackers which have successfully locked on targets, while the thin blue ones mean they are inactive and still roaming around to search for any potentially new targets. Labels are also displayed to help identify each tracker uniquely.

Some selected tracking results of the proposed approach are demonstrated in Figure 4 (Inactive trackers are not displayed for better illustration). The present soccer players are successfully tracked with uniquely assigned identifiers even under the severe interactions and occlusions. The involved collaborations among targets are depicted by the blue links representing the edges in the underlying Markov network structure (Please note that the network structure changes with time as shown in the Figure). With the help of collaborations among targets, the coalescence problem is successfully handled along the whole sequence.

In comparison, 16 multiple independent trackers M.i.T. are also tested to perform detecting and tracking with the same sequence. Every setting is the same as above except the missing of collaborative message passing. The comparison results using the proposed approach and M.i.T. are demonstrated in Figure 5, where the left column in the Figure is the original source frames, the middle column corresponds to our proposed approach, and the right one is the results from M.i.T.. For clear illustrations, only the tracking results from the pink areas of the original frames are shown in the middle and right columns. The frame numbers are 141 (top), 215 (bottom) respectively. In the following, for clarification purpose, all the identifiers we described are corresponding to the players in the proposed approach, i.e. the identifiers in the middle column, since there identities are maintained correctly for each player. In frame 141, the M.i.T. loses tracking the red team player 9 due to his previous crossing with the player 8. Actually, this interaction between these two players can be clearly seen in the frame 127 of Figure 4. In frame 215, the coalescence problem becomes more severe in the M.i.T. case, where the player 9, although previously lost in frame 141, and then sensed and tracked by other nearby inactive trackers, also "hijack" the tracker of the player 8, which leads to the lose track of player 8. In M.i.T., although the set of trackers equipped with the object detector may successfully cover all appearing targets within the dynamic scene, the inevitable coalescence problem there will result in targets identity switching, then dramatically hurt the tracking performance.

Secondly, a video sequence captured from a hockey game is tested, in which many hockey players appear and disappear in the field and present severe interactions. The tracking results of the sequence are originally reported in [12], which therefore provides a direct comparison of our algorithms with theirs. By explicitly introducing the collaborative mechanisms among the spatial adjacent trackers, our proposed approach robustly follows most of the hockey players and handles the coalescence problem satisfactorily, as can be seen in Figure 6, while in [12], a simple K-means clustering is proposed to maintain multiple modalities of the underlying particle filtering, therefore may easily result in the identity confusions of hockey players before and after clustering. Please note that the few hockey players are not successfully sensed in the sequence, which are mainly due to the imperfect object detector since it is only trained based the labelled data from 50 frames of the sequence.

5 Discussion and Conclusions

We propose a novel decentralized approach to multiple target tracking problem, where a set of autonomous while collaborative trackers are introduced to overcome the tracker coalescence problem with linear complexity. The autonomous means each individual tracker is self-aware since it can determine its own status, such as following a target or sensing potential new targets by an entropy-based evaluator; while the collaborative implies the set of trackers may also need to communicate with their neighboring ones to deal with the coalescence problem corporately. The theoretical foundation of this new approach is based on the variational analysis of a Markov network which reveals an intrinsically parallel variational message passing mechanism. In addition, a trained object detector is incorporated to help sense the potential newly appearing targets in the dynamic scene, therefore background subtraction is not necessary to

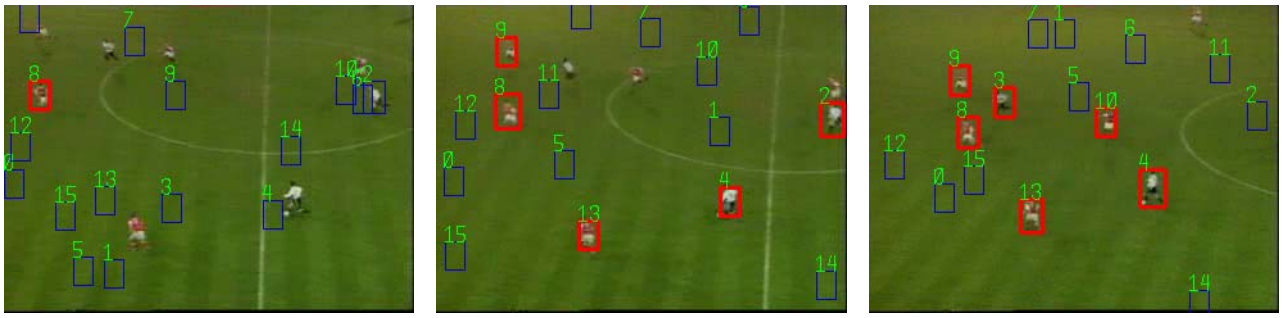

Figure 3: Soccer player detections using 16 autonomous trackers with local range AdaBoost detector, frame numbers are 2, 18, 29 respectively. The red thick rectangles illustrate active trackers, while the thin blue means inactive ones. See text for details.

.

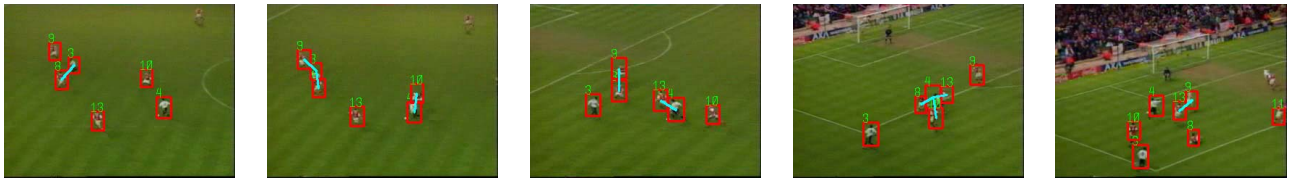

Figure 4: Tracking soccer players using the proposed approach, frame number 59, 75, 127, 186, 233. The blue links among the targets illustrate the structure of the Markov network. Please see the attached video for details.

our method. The use of object detectors within each tracker also supports the construction of an effective importance function, which leads to an more effective variational inference.

Since the proposed approach of collaborative tracking multiple targets is a general framework, it does not make any assumptions about the individual autonomous tracker. Therefore we are expecting to incorporate any promising progresses from the robust single target tracking methods into our formulation. One of the representatives is on-line feature selection in [1], where by exploiting the possible disjoint set of discriminative features for multiple targets when they are spatially far away, then when they are coming close, by constraining the corresponding trackers only employ the discovered disjoint feature set, the switching identity problem may be more reliably solved.

Acknowledgments

This work was supported in part by National Science Foundation Grants IIS-0347877, IIS-0308222, Northwestern faculty startup funds and Murphy Fellowship.

References

[1] R. T. Collins and Y. X. Liu. On-line selection of discriminative tracking features. In *Proc. IEEE Int'l Conf. on Computer Vision*, Nice, France, 2003.

[2] I. J. Cox and S.L. Hingorani. An efficient implementation of reid's multiple hypotheses tracking algorihm and its evaluation for the purpose of visual tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(2):138–150, 1996.

[3] W. Freeman, E. Pasztor and O. Carmichael. Learning low-level vision. *Int'l Journal of Computer Vision*, 40:25–47, 2000.

[4] I. Haritaoglu, D. Harwood and L. Davis. W4: Who? when? where? what? a real time system for detecting and tracking people. In *Proc. IEEE Int'l Conf. on Face and Gesture Recognition*, Nara, Japan, April 1998.

[5] M. Han, W. Xu, H. Tao and Y.H. Gong. An algorithm for multiple target trajectory tracking. In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, Washington, D.C., June 2004.

[6] C. Hue, J. Cadre and P. Perez. Tracking multiple targets with particle filtering. *IEEE Transactions on Aerospace and Electronic Systems*, 38(3):791–812, 2002.

[7] M. Isard and J. MacCormick. BraMBLe: A bayesian multiple-blob tracker. In *Proc. IEEE Int'l Conf. on Computer Vision*, pages 34–41, Vancouver, Canada, 2001.

[8] M. Jordan, Z. Ghahramani, T. Jaakkola and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 2000.

[9] Z. Khan, T. Balch and F. Dellaert. An MCMC-based particle filter for tracking multiple interacting targets. In *Proc. of European Conf. on Computer Vision*, 2004.

[10] D. Li, K. Wong, Y. Hu and A. Sayeed. Detection, classification and tracking of targets in distributed sensor networks. *IEEE Signal Processing Magazine*, 19(2), March, 2002.

[11] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple targets. In *Proc. IEEE Int'l Conf. on Computer Vision*, pages 572–578, Greece, 1999.

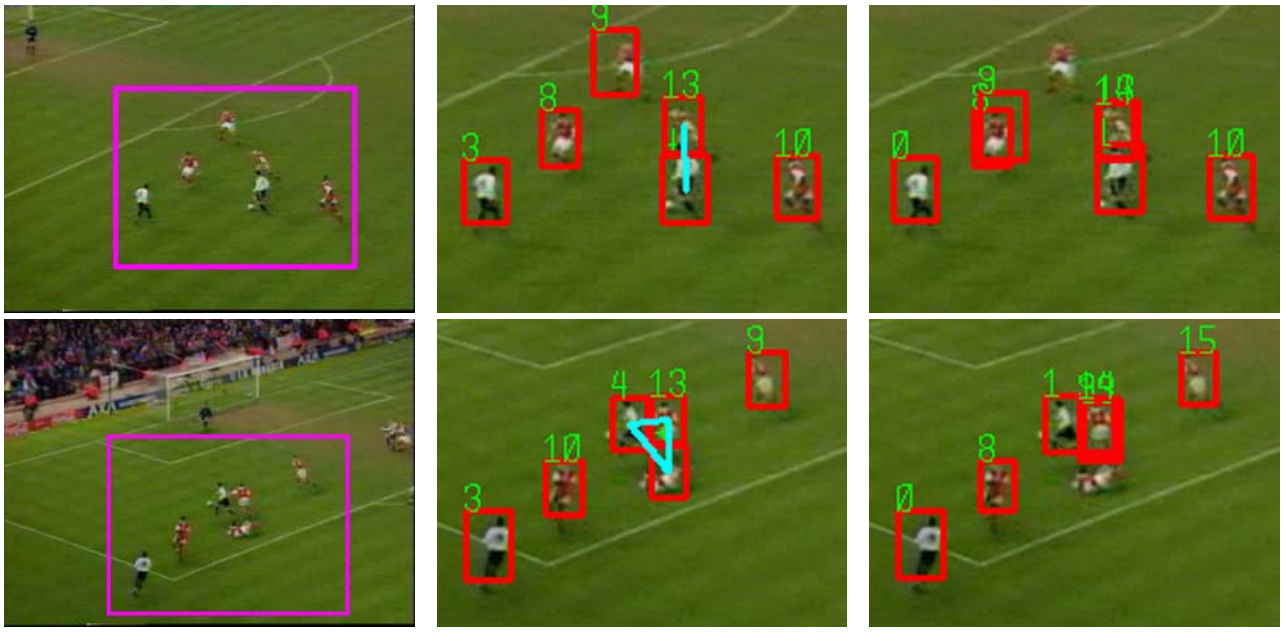

Figure 5: The comparison results of tracking soccer players using the proposed approach (middle column) and M.i.T. (right column). Left column is the corresponding source frames, where the pink areas are the actual showing regions in the middle and on the right for better illustration. Frame numbers are 141 (top), 215 (bottom). The blue links among the targets in the middle column illustrate the structure of the Markov network. See text for details.

Figure 6: Tracking hockey players with the proposed approach, frame number 31, 39, 63, 64, 115. The blue links among the targets illustrate the structure of the Markov network. Please see the attached video for details. The authors acknowledge Mr. Kenji Okuma for providing the test data on the website.

[12] K. Okuma, A. Taleghani, N. D. Freitas, J. J. Little and D. G. Lowe. A boosted particle filter: multitarget detection and tracking. In *Proc. of European Conf. on Computer Vision*, 2004.

[13] C. Rasmussen and G. Hager. Probabilistic data association methods for tracking complex visual targets. *IEEE T-PAMI*, pages 560–576, Jun. 2001.

[14] Y. Rui and Y. Q. Chen. Better proposal distributions: target tracking using unscented particle filter. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Kauai, Hawaii, 2001.

[15] E. Sudderth, M. Mandel, W. Freeman and A. Willsky. Distributed Occlusion Reasoning for Tracking with Nonparametric Belief Propagation. In *Proc. Neural Information Processing Systems*, June 2004.

[16] H. Tao, H. Sawhney and R. Kumar. A sampling algorithm for detecting and tracking multiple targets. In *Proc. ICCV'99 Workshop on Vision Algorithm*, Corfu, Greece, 1999.

[17] J. Vermaak, A. Doucet and P. Perez. Maintaining multimodality through mixture tracking. In *Proc. IEEE Int'l Conf. on Computer Vision*, 2003.

[18] P. Viola and M. Jones. Rapid target detection using a boosted cascade of simple features. In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, 2001.

[19] Y. Wu, G. Hua and T. Yu. Tracking articulated body by dynamic Markov network. In *Proc. IEEE Int'l Conf. on Computer Vision*, pages 1094–1101, Nice, France, Oct. 2003.

[20] T. Yu and Y. Wu. Collaborative tracking of multiple targets. In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, Washington, D.C., June 2004.

[21] T. Zhao and R. Nevatia. Tracking Multiple Humans in Crowded Environment. In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, Washington, D.C., June 2004.