Multi-object Tracking Under Occlusion Using Dual-Mode Graph Embedding

Jin Gao, Junliang Xing, Weiming Hu, Xiaoqin Zhang and Ruiguang Hu
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
Beijing, China
E-mail:{jgao10|jlxing|wmhu|xqzhang|rghu}@nlpr.ia.ac.cn

***Abstract*—In this paper we address the problem of tracking multiple objects to know how objects are moving (*e.g.* occlusion relationships) while interacting with each other in a group, given models of their appearances that are learned online even when occlusion occurs. This aim is very different from the recently popular detection-based tracklets association approaches. In our approach, occlusion relationships between multiple objects are explicitly defined and deduction of the occlusion relationships is integrated into the whole tracking framework. Specifically, we deduce the joint state estimation problem in the multi-object tracking in a new decentralized strategy, that the single object tracking and the multi-object separating are viewed as one-versus-rest classification problems based on graph embedding framework. Two kinds of discriminative subspaces are learned: one for single object tracking which is robust to various appearance variations; the other for occlusion reasoning and decentralizing. Partial disappearance can also be addressed as an occlusion problem by this strategy. Experimental results demonstrate the effectiveness of our method.**

***Keywords*-multi-object tracking; occlusion reasoning; partial occlusion and disappearance;**

I. Introduction

Visual surveillance systems are required to keep track of targets as they move through the scene even when they are occluded by or interacting with other objects in a group. It is desirable for visual surveillance systems to know how objects are moving while interacting with each other. The problem that we address in this paper is how to track multiple objects and reason their occlusion relationships simultaneously while they are interacting with each other in a group, given models of their appearances that are learned online. This problem is also important for other video analysis applications such as video retrieval and video archival. The aim of this paper is very different from the recently popular detection-based tracklets association approaches (see below).

Many detection-based tracking methods have been proposed for multi-pedestrian tracking recently [1], [2], [3], [4], [5], [6]. These off-line methods first detect the pedestrians by a pre-trained detector and then assign the detection responses to the tracked trajectories using a variety of data association strategies. The performance of these methods greatly depends on the accuracy of pedestrian detection. These methods also can not reason the occlusion relationships when occlusion occurs. Despite of the effectiveness of these methods in occlusion scenario, it is still a challenge to track multiple objects (not only pedestrians) in a general way without prior knowledge about objects, but with online learned object appearance model.

There has been much work on tracking multiple objects using object appearance model. Generally speaking, the motions of multiple objects have to be jointly estimated from the mixed visual observations when occlusion occurs. Some existing methods (*e.g.* [7], [8]) concatenate the states of different objects in a centralized fashion, view the multi-object tracking as a joint state estimation problem and search a rather high dimensional solution space. In this paper, we propose a new strategy to decentralize the joint tracker into discriminative appearance model based individual trackers. Specifically, we learn two kinds of discriminative subspaces based on graph embedding framework: 1) one kind for individually tracking when the joint tracker is decentralized, which can make full use of the information in the background and is robust to various appearance variations including occlusions; 2) the other kind for occlusion reasoning which is used to decentralize the joint tracker (see Sec. III-A). We call the first kind as single object tracking subspace (SOTS), and the other kind as multi-object discriminant subspace (MODS). It is noted that, this paper concentrates more on the severe occlusions and occlusion relationships among tracked objects. Our individual tracker based on SOTS is robust to non-severe occlusions, such as object self-occlusion, and occlusion by other scene objects. Partial disappearance can be addressed as an occlusion problem in our proposed tracking framework.

Some existing methods (*e.g.* [9], [10], [11]) also decentralize the joint tracker using different strategies. Hu *et al*. [9] adopt a selective updating and matching strategy based on block-division. Zhang *et al*. [11] introduce the species concept into the PSO framework and the occlusion between different objects is modeled as species competition. Thus the joint tracker can be decentralized into individual trackers, each of which try to maximize its own visual evidence. These two studies can reason the occlusion relationship, however they both adopt the incremental subspace learning based generative appearance model, which discards the

978-1-4799-2190-4/13 $31.00 © 2013 IEEE
DOI 10.1109/ACPR.2013.77

information for classification in the background leading to tracking degradations. Yang *et al*. [10] decentralize the joint tracker into a set of simple individual target trackers, each of which estimates its own state in the Nash Equilibrium of a game when the objects are in close vicinity. However, its individual trackers are based on the mean-shift tracker, which may not guarantee robust tracking even when occlusion does not occur.

Some methods (*e.g.* [12], [13]) also concentrate more on the same aim as ours: tracking multiple objects to know how objects are moving (*e.g.* occlusion relationships) while interacting with each other in a group. However, these two methods are specially designed for tracking multiple people or pedestrians, while not for more general objects (*e.g.* faces, pedestrians, cars, and so on). In addition, they both use the spatial-color mixture of Gaussians appearance model, which needs the color information of the videos. Their methods are not suitable for dealing with the gray-scale videos.

The rest of the paper is organized as follows. Sec. II details the methodology of our designed graph embedding based discriminative learning. In Sec. III, we introduce our multi-object tracking framework. Experimental results are reported in Sec. IV. Finally, we conclude the paper in Sec. V.

II. Discriminative Learning Methodology

Based on the graph embedding framework [14], [15], we design our discriminative learning based methodology for tracking, which imposes the embedding space to have two different properties with others: 1) two *nearby* samples (not all the samples) of the same class stay close to one another; 2) two samples of different classes stay far apart.

Let $\mathbf{x}_i \in \mathbb{R}^D (i = 1, 2, \ldots, l)$ be D-dimensional vectors to represent the graph vertices corresponding to the labeled samples, and $y_i \in \{1, 2, \ldots, C\}$ be associated class labels, where l is the number of labeled samples and C is the number of classes. Let n_c be the number of samples in the class c: $\sum_{c=1}^{C} n_c = l$. Using the information of the labeled samples, we aim to find a discriminative embedding space and map $\mathbf{X} \equiv (\mathbf{x}_1|\mathbf{x}_2|\cdots|\mathbf{x}_l) \in \mathbb{R}^{D\times l} \mapsto \mathbf{Z} \equiv (\mathbf{z}_1|\mathbf{z}_2|\cdots|\mathbf{z}_l) \in \mathbb{R}^{R\times l}$, where $R < D$ and is the dimension of the embedding space, such that in the embedding space unlabeled samples are easy to be labeled by a simple classifier (*i.e.* the nearest neighbor algorithm). To achieve this goal, we need to construct two graphs: the intra-class compactness graph $\mathcal{G} = \{\mathbf{X}, \mathbf{W}\}$ and the inter-class separability graph $\mathcal{G}^p = \{\mathbf{X}, \mathbf{W}^p\}$, where $\mathbf{W}$ and $\mathbf{W}^p$ are edge weight matrices.

We assume the linear projection in the linear extension of graph embedding framework as $\mathbf{Z} = \mathbf{P}^T\mathbf{X}$, where $\mathbf{P}$ is a $D \times R$ transformation matrix. Thus, the intra-class compactness is characterized by

$$\tilde{S} = \sum_{i,j} \|\mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j\|^2 w_{ij} = 2\,\mathrm{tr}\left(\mathbf{P}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P}\right) \ , \quad (1)$$

and the inter-class separability is characterized by

$$\tilde{S}^p = \sum_{i,j} \|\mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j\|^2 w^p_{ij} = 2\,\mathrm{tr}\left(\mathbf{P}^T\mathbf{X}\mathbf{L}^p\mathbf{X}^T\mathbf{P}\right) \ , \quad (2)$$

where, the Laplacian matrices $\mathbf{L}$ and $\mathbf{L}^p$ of $\mathcal{G}$ and $\mathcal{G}^p$ are defined by the diagonal matrices $\mathbf{D}$ and $\mathbf{D}^p$ as:

$$\mathbf{L} = \mathbf{D} - \mathbf{W}, \quad D_{ii} = \sum_{j\neq i} w_{ij}, \quad \forall i \quad (3)$$

$$\mathbf{L}^p = \mathbf{D}^p - \mathbf{W}^p, \quad D^p_{ii} = \sum_{j\neq i} w^p_{ij}, \quad \forall i \ . \quad (4)$$

Further more, we add the edges between any vertex pair in $\mathcal{G}$ and $\mathcal{G}^p$ by the local scaling method [16] as follows:

$$\begin{cases} w_{ij} = A_{ij}/n_c, w^p_{ij} = A_{ij}(1/l - 1/n_c), & \text{if } y_i = y_j, \\ w_{ij} = 0, w^p_{ij} = 1/l, & \text{otherwise,} \end{cases} \quad (5)$$

where

$$A_{ij} = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(\sigma_i\sigma_j)\right), \quad (6)$$

$\sigma_i = \left\|\mathbf{x}_i - \mathbf{x}_i^{(k)}\right\|$, and $\mathbf{x}_i^{(k)}$ is the kth nearest neighbor in the same class of the sample $\mathbf{x}_i$. k is empirically chosen as 7 based on [16].

Our discriminative learning analysis aims at finding the optimal projection direction that optimizes the graph-preserving criterion

$$\mathbf{P}^* = \underset{\mathbf{P}\in R^{D\times R}}{\mathrm{argmin}}\ \mathrm{tr}\left(\frac{\mathbf{P}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P}}{\mathbf{P}^T\mathbf{X}\mathbf{L}^p\mathbf{X}^T\mathbf{P}}\right) \ , \quad (7)$$

where the analytic form of $\mathbf{P}^*$ is obtained by solving a generalized eigenvalue problem as follows:

$$\mathbf{P}^T\mathbf{X}\mathbf{L}^p\mathbf{X}^T\mathbf{P}\varphi = \lambda\mathbf{P}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P}\varphi \ . \quad (8)$$

Denoting R principal generalized eigenvectors corresponding to the R largest eigenvalues of Eq. (8) as $\{\varphi_r\}_{r=1}^R$, we can obtain the discriminative projection $\mathbf{P}^* = (\varphi_1|\varphi_2|\cdots|\varphi_R)$.

Kernel trick is widely used to enhance the separability of the linear discriminative learning leading to the non-linear extension of graph embedding. Let $\phi : \mathbf{x} \mapsto \mathcal{H}$ be a function mapping the points in the input space to a high-dimensional Hilbert space. For a proper chosen ϕ, we replace the explicit mapping with the inner product $K(\mathbf{x}_i, \mathbf{x}_j) = \langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle$. Here we use Gaussian kernel to define this product: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2\right)$, where σ is set to the average pairwise distances among all the data points. For convenience, we rewrite the vertex matrix in the Hilbert space as $\mathbf{X}^\phi \equiv (\phi(\mathbf{x}_1)|\phi(\mathbf{x}_2)|\cdots|\phi(\mathbf{x}_l))$. Then, we can rewrite Eq. (7) as follows:

$$\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha}}{\mathrm{argmin}}\ \mathrm{tr}\left(\frac{\boldsymbol{\alpha}^T\mathbf{K}\mathbf{L}\mathbf{K}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\mathbf{K}\mathbf{L}^p\mathbf{K}\boldsymbol{\alpha}}\right) \quad (9)$$

where $\mathbf{K} = \mathbf{X}^{\phi T}\mathbf{X}^\phi$. A data point in the Hilbert space can be embedded into R-dimensional subspace by: $\phi(\mathbf{x}) \mapsto \mathbf{z} = \boldsymbol{\alpha}^{*T}K(:, \mathbf{x})$, where $K(:, \mathbf{x}) = (K(\mathbf{x}_1, \mathbf{x})|\cdots|K(\mathbf{x}_l, \mathbf{x}))^T$.

Algorithm 1 Occlusion Reasoning

Input: Current positive sample buffers $\mathcal{B}_f^i$, tracking results s_{t,k_i}^*, optimal observations o_{t,k_i}, where $i = 1, 2$.

Output: The occlusion relationship $R = \phi(k_2) - \phi(k_1)$, where $\phi(k_i) = 1$ indicates k_i is occluded, and 0 indicates not; updated $\mathcal{B}_f^i$.

1: **if** overlap$(s_{t,k_1}^*, s_{t,k_2}^*) > Th_1$ **then**
2: For each sample in $\mathcal{B}_f^i$, extract feature vector $\mathbf{x}_j^i$, where $j = 1, \ldots, |\mathcal{B}_f^i|$, and its label is $y_j^i = i$;
3: Learn MODS by Eq. (9) using $\{\mathbf{x}_j^i, y_j^i\}_{j=1,\ldots,|\mathcal{B}_f^i|}^{i=1,2}$;
4: Extract feature vectors of o_{t,k_i} and embed them into MODS as $\mathbf{z}^i$;
5: Conduct outlier detection in MODS: if $\mathbf{z}^i$ is an outlier as for $\{\mathbf{z}_j^i\}_{j=1,\ldots,|\mathcal{B}_f^i|}$, $\phi(k_i) = 1$; else, $\phi(k_i) = 0$;
6: If $R = 0$, update buffers $\mathcal{B}_f^i$ using the pseudo-object observations o'_{t,k_i}; else, update $\mathcal{B}_f^i$ of the occluded object k_i ($\phi(k_i) = 1$) using o'_{t,k_i} and update $\mathcal{B}_f^i$ of the other using o_{t,k_i}.
7: **end if**

III. Multi-object Tracking Framework

In this section, we introduce our dual-mode graph embedding model for decentralizing the joint tracker and tracking each object individually based on SOTS and MODS.

A. Joint Likelihood Maximization

Denote the state of the kth object by $s_{t,k}$. Its corresponding support is denoted by $o_{t,k}$, *i.e.* the set of pixels within the region of it. Thus, the states of a number of M objects can be estimated by maximizing the joint likelihood

$$\mathcal{S}_t^* = \underset{s_{t,1},\ldots,s_{t,M}}{\operatorname{argmax}} P\left(\bigcup_{k=1}^{M} o_{t,k} | s_{t,1}, \ldots, s_{t,M}\right), \quad (10)$$

where $\mathcal{S}_t = \{s_{t,k}, k = 1, \ldots, M\}$, and t represents the tth image frame. If no occlusion is present, the above joint optimization can be done by maximizing the individual observation likelihood independently:

$$s_{t,k}^* = \underset{s_{t,k}}{\operatorname{argmax}} P\left(o_{t,k} | s_{t,k}\right), \ k = 1, \ldots, M. \quad (11)$$

If occlusion happens between objects k_1 and k_2, as shown in Fig. 1(a), we divide the observation of each object k_i into two parts: non-overlapping part $\tilde{o}_{t,k_i}$ and overlapping part $\hat{o}_{t,k_i}$ (see Fig. 1(b)). We substitute $\hat{o}_{t,k_i}$ with the corresponding part of k_i's mean observation (referred to [17]), and thus generate a pseudo-object observation o'_{t,k_i} (see Fig. 1(c)). When we have reasoned the occlusion relationship between these two objects (see Sec. III-B), the tracking problem of these two objects hence can be formulated as follows (without loss of generality, here we assume object k_2 is occluded):

$$s_{t,k_1}^* = \underset{s_{t,k_1}}{\operatorname{argmax}} P\left(o_{t,k_1} | s_{t,k_1}\right) \quad (12)$$

$$s_{t,k_2}^* = \underset{s_{t,k_2}}{\operatorname{argmax}} P\left(o'_{t,k_2} | s_{t,k_2}, s_{t,k_1}^*\right). \quad (13)$$

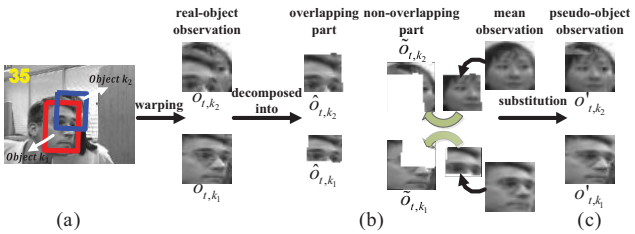


Figure 1: Observation decomposition and substitution of two objects under occlusion. (a) shows a frame in the *Girl* sequence. The observation decomposition is exhibited in (b). (c) shows the pseudo-object observations after substitution.

Note that the tracking of the occluded object relies on the tracking accuracy of the non-occluded object. In the next, we will show how to perform occlusion reasoning to find the non-occluded object and use it to track the occluded one.

Algorithm 2 Individually Tracking Under Occlusion

Input: Previous positive sample buffers $\mathcal{B}_f^i$, negative sample buffers $\mathcal{B}_b^i$, the tracking results s_{t-1,k_i}^* in the frame $t-1$, current sampled candidate object states $\{s_{t,k_i}^m\}_{m=1}^{\mathcal{M}}$ using the particle filters.

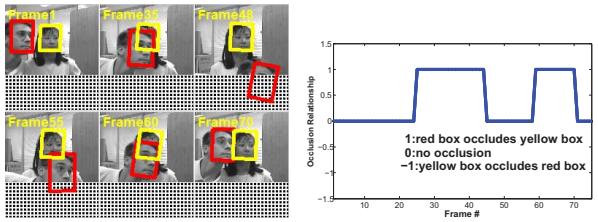
Output: Current tracking results s_{t,k_i}^*, updated $\mathcal{B}_b^i$.

1: For each sample in $\mathcal{B}^i = \mathcal{B}_f^i \bigcup \mathcal{B}_b^i$, extract feature vector $\mathbf{x}_j^i$, and set labels y_j^i of samples in $\mathcal{B}_f^i$ as 1 and in $\mathcal{B}_b^i$ as 2, where $j = 1, \ldots, |\mathcal{B}^i|$;
2: Learn SOTS for k_i by Eq. (9) using $\{\mathbf{x}_j^i, y_j^i\}_{j=1,\ldots,|\mathcal{B}^i|}$;
3: **if** overlap$(s_{t-1,k_1}^*, s_{t-1,k_2}^*) > Th_1$ **then**
4: **if** $R = 0$ or $\phi(k_i) = 0$ **then**
5: Adopt strategy 1 in Sec. III-C to estimate s_{t,k_i}^* and update $\mathcal{B}_b^i$;
6: **else if** $\phi(k_i) = 1$ **then**
7: Adopt strategy 2 in Sec. III-C to estimate s_{t,k_i}^* and update $\mathcal{B}_b^i$.
8: **end if**
9: **end if**

B. Occlusion Reasoning

We learn MODS for occlusion reasoning based on positive sample buffers, each of which consists of foreground samples of related object. The samples are the optimal observations corresponding to the tracking results of the previous frames. Additionally, block-division based representation method has achieved good results in occlusion reasoning [9], so we also represent each observation as a block-division based feature vector. Specifically, we extract covariance matrix descriptors for all 4×4 cells, and represent each cell as a vector generated by Log-Euclidean mapping and unfolding (referred to [9]), resulting in a feature vector. It is also suitable for approximating the probabilities for the observations in Sec. III-C.

Without loss of simplicity, we only consider the occlusion between objects k_1 and k_2. The occlusion between three or more objects can be considered similarly. Firstly, we use Overlap-Criterion $\text{overlap}(s_{t,k_1}^*, s_{t,k_2}^*) = \frac{\text{area}(s_{t,k_1}^* \bigcap s_{t,k_2}^*)}{\text{area}(s_{t,k_1}^* \bigcup s_{t,k_2}^*)}$ to detect wether occlusion happens or not between k_1



(a) (b)

Figure 2: Tracking under disappearance and occlusion with short term in the video *Girl*.

and k_2. Secondly, if occlusion happens, we learn MODS based on discriminative learning methodology, and then map current optimal observations o_{t,k_i} into MODS as $\mathbf{z}^i$. Lastly, we conduct outlier detection in MODS for reasoning the relationship between k_1 and k_2. If $\mathbf{z}^i$ is an outlier as for k_i's positive samples in MODS, object k_i is occluded, and vice versa. Algorithm 1 details the reasoning procedure.

C. Individually Tracking Under Occlusion

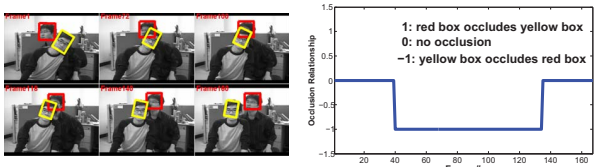
We learn SOTS for k_i based on its aforementioned positive sample buffer and negative sample buffer which consists of its bad observations collected from the previous frames. When SOTS is learned and characterized by $\mathbf{z}_+^i$, $\boldsymbol{\alpha}^i$ and $K^i(:,\cdot)$, where $\mathbf{z}_+^i$ is the center of the positive samples in SOTS, the probability that a feature vector $\mathbf{x}$ is generated from the distribution of k_i's positive samples can be measured by:

$$P(\mathbf{x}|\mathbf{z}_+^i, \boldsymbol{\alpha}^i) \propto \exp(-\|\mathbf{z}_+^i - \boldsymbol{\alpha}^{iT} K^i(:,\mathbf{x})\|). \quad (14)$$

We consider the individually tracking under the occlusion between two objects k_i and $k_{\bar{i}}$. If $i = 1$, $\bar{i} = 2$, and vice versa. The individually tracking under the non-occlusion scenario can be considered similarly. We use the particle filters (referred to [17]) to sample candidate object states $\{s_{t,k_i}^m\}_{m=1}^{\mathcal{M}}$. We adopt two strategies for individually tracking when occlusion occurs:

1) Get candidate object observations $\{o_{t,k_i}^m\}_{m=1}^{\mathcal{M}}$ corresponding to $\{s_{t,k_i}^m\}_{m=1}^{\mathcal{M}}$, extract feature vectors of them $\{\mathbf{x}_{t,m}^i\}_{m=1}^{\mathcal{M}}$, let $P(o_{t,k_i}^m|s_{t,k_i}^m) = P(\mathbf{x}_{t,m}^i|\mathbf{z}_+^i, \boldsymbol{\alpha}^i)$, determine optimal object states s_{t,k_i}^*, update $\mathcal{B}_b^i$ using the bad observations;
2) Get candidate pseudo-object observations $\{(o_{t,k_i}^m)'\}_{m=1}^{\mathcal{M}}$ corresponding to $\{s_{t,k_i}^m\}_{m=1}^{\mathcal{M}}$, extract feature vectors of them $\{\mathbf{x}_{t,m}^i\}_{m=1}^{\mathcal{M}}$, let $P((o_{t,k_i}^m)'|s_{t,k_i}^m, s_{t,k_{\bar{i}}}^*) = P(\mathbf{x}_{t,m}^i|\mathbf{z}_+^i, \boldsymbol{\alpha}^i)$, determine optimal object states s_{t,k_i}^*, update $\mathcal{B}_b^i$ using the bad pseudo-object observations.

Algorithm 2 details the tracking procedure. It is noted that the optimal real-object observation o_{t,k_i} corresponding to s_{t,k_i}^* should be as different from object $k_{\bar{i}}$ as possible, so that merging can be avoided. This can be done by adding a penalty term to $P((o_{t,k_i}^m)'|s_{t,k_i}^m, s_{t,k_{\bar{i}}}^*)$.



(a) (b)

Figure 3: Tracking under occlusion with long term in the video *TwoFaces*.

IV. Experimental Results

In this section, we design experiments to demonstrate the superior properties of our approach. Firstly, two videos (*Girl* [9], *TwoFaces* [9]) are used to demonstrate the effectiveness of our novel strategy on three aspects: tracking under occlusion, reasoning occlusion relationship, and handling partial disappearance. Note that the success of tracking and occlusion reasoning in this part is checked by our own judgment. Secondly, we compare our approach with two related methods [9], [10] qualitatively and quantitatively in other two publicly available videos (*ThreePastShop2cor*, *PeopleVehicle*), which prove that our strategy for decentralizing the joint tracker could handle merging and splitting well. The video *ThreePastShop2cor* is taken from the CAVIAR Test Case Scenarios dataset. [1] This video captures people moving in a shopping center with frequent occlusions and interactions. The video *PeopleVehicle* is the first view of the Dataset 1 sequence from the PETS 2001 benchmark. [2] This video captures outdoor people and vehicle moving.

A. Effectiveness of Occlusion Handling

The results of occlusion reasoning are illustrated using the recovered occlusion relationship diagram whose x-coordinate is the frame number, and y-coordinate is the occlusion relationship. The video *Girl* shows a man (red) occludes a woman (yellow) twice and he also partially disappears from the scene and then reappears (see Fig. 2(a)). Fig. 2(b) shows occlusion relationship between them recovered from the video *Girl*. Note that the woman undergos occlusion with short term (less than 30 frames) every time. Additionally, we address the man's disappearance as an occlusion problem, that he is occluded by a hypothetical object (black-and-white grid in Fig. 2(a)). The video *TwoFaces* is used to demonstrate the effectiveness when object undergos occlusion with long term (more than 80 frames). Fig. 3(a) shows some tracking results of this video and Fig. 3(b) shows the recovered occlusion relationship.

B. Comparison with Other Methods

To show the superiority of our approach over Yang's work [10] and Hu's work [9], we perform experiments in other two publicly available videos (see Fig. 4 and Fig. 5).

[1] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[2] http://www.hitech-projects.com/euprojects/cantata/datasets_cantata/dataset.html.

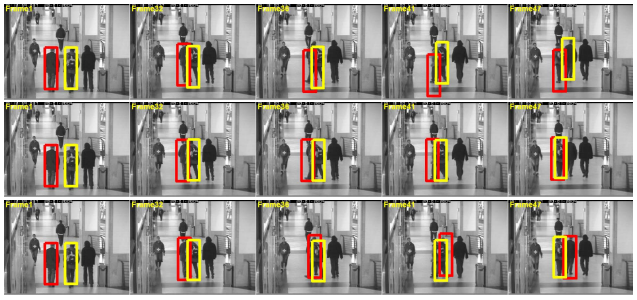


Figure 4: Qualitative results in the video *ThreePastShop2cor*.

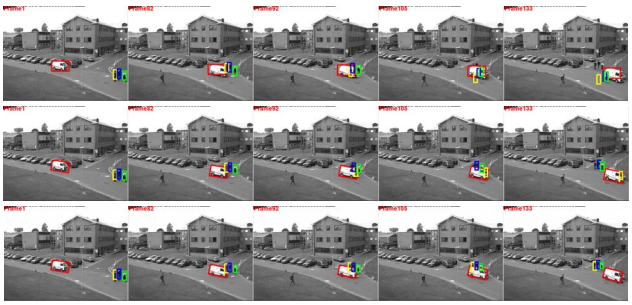


Figure 5: Qualitative results in the video *PeopleVehicle*.

The top row is Yang's work, middle row is Hu's work and bottom row is our approach. Quantitative evaluation is conducted in following aspects: STF (successfully tracked frames) and ACLE (average center location errors) between the estimated position and the groundtruth. Table I shows the quantitative comparison.

V. Conclusion

In this paper, we have proposed a new strategy to decentralize the joint tracker for tracking multiple objects to know how objects are moving (*e.g.* occlusion relationships) while interacting with each other in a group based on a dual-mode graph embedding model. Experimental results demonstrate that the proposed approach can effectively and accurately track objects under occlusion and partial disappearance, obtaining correct occlusion relationship.

Acknowledgment

This work is partly supported by NSFC (Grant No. 60935002), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504), the Natural Science Foundation of Beijing (Grant No. 4121003), and The Project Supported by Guangdong Natural Science Foundation (Grant No. S2012020011081).

Table I: Quantitative results for the other two public videos.

Methods	Yang's Work		Hu's Work		Our Approach	
Evaluation	**ACLE**	**STF**	**ACLE**	**STF**	**ACLE**	**STF**
ThreePastShop2cor A(Red)	26.2	34/62	21.8	34/62	6.8	56/62
ThreePastShop2cor B(Yellow)	23.6	39/62	4.4	61/62	4.7	62/62
PeopleVehicle A(Red)	16.7	141/141	5.6	141/141	4.3	141/141
PeopleVehicle B(Yellow)	31.9	88/141	47.9	85/141	3.9	140/141
PeopleVehicle C(Blue)	23.5	93/141	2.7	141/141	2.6	141/141
PeopleVehicle D(Green)	25.8	86/141	2.9	140/141	1.6	141/141

References

[1] A. Ess, B. Leibe, K. Schindler, and L. Gool, "A mobile vision system for robust multi-person tracking," in *Proc. CVPR*, 2008.

[2] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *Proc. ECCV*, 2008.

[3] D. Mitzel, E. Horbert, A. Ess, and B. Leibe, "Multi-person tracking with sparse detection and continuous segmentation," in *Proc. ECCV*, 2010.

[4] B. Yang and R. Nevatia, "An online learned crf model for multi-target tracking," in *Proc. CVPR*, 2012.

[5] A. Andriyenko, K. Schindler, and S. Roth, "Discrete-continuous optimization for multi-target tracking," in *Proc. CVPR*, 2012.

[6] S. Tang, M. Andriluka, and B. Schiele, "Detection and tracking of occluded people," in *Proc. BMVC*, 2012.

[7] Z. Khan, T. Balch, and F. Dellaert, "Mcmc-based particle filtering for tracking a variable number of interacting targets," *IEEE Trans. PAMI*, vol. 27, no. 11, pp. 1805–1819, 2005.

[8] T. Zhao and R. Nevatia, "Tracking multiple human in crowded environment," in *Proc. CVPR*, 2004.

[9] W. Hu, X. Li, W. Luo, X. Zhang, S. Maybank, and Z. Zhang, "Single and multiple object tracking using log-euclidean riemannian subspace and block-division appearance model," *IEEE Trans. PAMI*, vol. 34, no. 12, pp. 2420–2440, 2012.

[10] M. Yang, T. Yu, and Y. Wu, "Game-theoretic multiple target tracking," in *Proc. ICCV*, 2007.

[11] X. Zhang, W. Hu, W. Qu, and S. Maybank, "Multiple object tracking via species-based particle swarm optimization," *IEEE Trans. CSVT*, vol. 20, no. 11, pp. 1590–1602, 2010.

[12] A. M. Elgammal and L. S. Davis, "Probabilistic framework for segmenting people under occlusion," in *Proc. ICCV*, 2001.

[13] W. Hu, X. Zhou, M. Hu, and S. Maybank, "Occlusion reasoning for tracking multiple people," *IEEE Trans. CSVT*, vol. 19, no. 1, pp. 114–121, 2009.

[14] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. PAMI*, vol. 29, pp. 40–51, 2007.

[15] H. T. Chen, H. W. Chang, and T. L. Liu, "Local discriminant embedding and its variants," in *Proc. CVPR*, 2005.

[16] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Proc. NIPS*, 2005.

[17] D. Ross, J. Lim, R. Lin, and M. Yang, "Incremental learning for robust visual tracking," *Int. J. Comput. Vision*, vol. 77, no. 1, pp. 125–141, 2008.