Group Structure Preserving Pedestrian Tracking in a Multicamera Video Network

Zhixing Jin, Le An, and Bir Bhanu, *Fellow, IEEE*

***Abstract*— Pedestrian tracking in video has been a popular research topic with many practical applications. In order to improve tracking performance, many ideas have been proposed, among which the use of geometric information is one of the most popular directions in recent research. In this paper, we propose a novel multicamera pedestrian tracking framework, which incorporates the structural information of pedestrian groups in the crowd. In this framework, first, a new cross-camera model is proposed, which enables the fusion of the confidence information from all camera views. Second, the group structures on the ground plane provide extra constraints between pedestrians. Third, the structured support vector machine is adopted to update the cross-camera model for each pedestrian according to the most recent tracked location. The experiments and detailed analysis are conducted on challenging data. The results demonstrate that the improvement in tracking performance is significant when a group structure is integrated.**

***Index Terms*— Group structure, multicamera, pedestrian tracking.**

I. Introduction

PEDESTRIAN tracking is one of the most important topics in video technology that has drawn the attention of many researchers over the years. It has made crucial contributions to many important application areas, such as video monitoring, security, surveillance, and resource management [1]. During the past decades, there has been significant progress for pedestrian tracking, and researchers have switched their attention from simple to much more complex scenarios [2]–[9]. For other computer vision tasks, such as image retrieval, recent work [10] can jointly utilize both visual and textual information to achieve satisfactory performance on finding similar semantic content in complex scenes. However, confronting the task of pedestrian tracking, which mainly relies on visual information, an acceptable performance still has not been achieved due to various challenges encountered in the real world.

One of the aspects that defines the complexity of a scenario is the amount of occlusion among pedestrians in a video. For a complex scenario, the occlusions of pedestrians can be significant, which makes it one of the biggest challenges in pedestrian tracking. The occlusion for each pedestrian can be caused by various sources, such as static objects in surrounding environment (e.g., buildings and trees) and other pedestrians in the same scene, especially when pedestrians are in a crowd. When occlusions occur, the appearance and shape models, which are widely used in traditional tracking approaches, become less reliable, which leads to degradation in the performance.

To alleviate the adverse impact of occlusions, researchers have proposed various methods from different perspectives, including dividing pedestrian body into different parts and training separate models for them [3], [11], changing camera view angles to a bird view to avoid occlusions [12], using information from both past and future frames in a certain sliding window to construct pedestrian trajectories [3], [5], or deploying multiple cameras with overlapping field-of-views (FOVs) and integrating information from them [7], [13].

The fact that pedestrians have similar appearances is another challenge that can cause a significant drop in tracking performance. For different pedestrians, it is almost impossible to distinguish them from one another using only the shape features computed under normal camera resolution. In addition, the color features for different pedestrians may also be similar in complex scenarios, especially when there are many occlusions. Therefore, additional information needs to be incorporated when tracking pedestrians in complex scenarios. For example, the spatial and temporal information is one of the most useful candidates that can provide significant help. The use of the spatial and temporal information includes but is not limited to checking distances between detections from consecutive frames when connecting them to form tracklets [3], grouping pedestrians based on distance and velocity metrics and taking the advantage of group information in pedestrian tracking [5], and generating confidence masks according to pedestrian velocity information when associating detections and trackers [4].

In this paper, the pedestrian tracking problem that we focus on contains crowded scenes, which may have many occlusions when only a single camera is used. Therefore, multiple cameras with overlapping FOVs can be utilized in order to obtain better tracking performance. In addition, the spatial and

Manuscript received November 13, 2015; revised February 11, 2016; accepted April 20, 2016. Date of publication May 10, 2016; date of current version October 3, 2017. This work was supported in part by the National Science Foundation under Grant 1330110 and in part by the Office of Naval Research under Grant N00014-12-1-1026. The contents of the information does not necessarily reflect the position or policy of the U.S. Government. This paper was recommended by Associate Editor Y. Wu. *(Corresponding author: Le An.)*

Z. Jin is with the Department of Computer Science and Engineering, University of California Riverside, Riverside, CA 92521 USA (e-mail: jinz@cs.ucr.edu).

L. An is with the National Key Laboratory of Science and Technology on Multi-spectral Information Processing, School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: lan004@ucr.edu).

B. Bhanu is with the Center for Research in Intelligent Systems, University of California Riverside, Riverside, CA 92521 USA (e-mail: bhanu@cris.ucr.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2016.2565998

temporal relationships among pedestrians are explored using group structures. For each group, its structure is represented by a minimum spanning tree that connects pedestrians, and a structure preserving object tracking (SPOT) approach is adopted [14]. At each time step, a group-based tracking stage that can simultaneously locate groups and their pedestrians is used instead of the original individual-based tracking.

The framework for the proposed approach is shown in Fig. 1. For each pedestrian, a support vector machine (SVM) classifier is trained for each camera view during the initialization stage. At each time step, multiple groups are extracted from the whole crowd based on the locations and velocities of all pedestrians from the previous time step [5]. The structure for each group is computed as a minimum spanning tree that connects all the pedestrians [14]. Then, a combined confidence map for each group is computed on the ground plane based on the classifiers as well as the group structure. The group location is tracked using this confidence map, and the location for each pedestrian in the group is estimated at the same time. Finally, the estimated pedestrian location on the ground plane is used to update the cross-camera pedestrian model when the classification result has a high confidence. As demonstrated in recent research, the group information can be used to improve tracking performance [5], [14]. Therefore, the proposed framework is expected to outperform the tracking approaches that do not exploit the structural information for each group.

The rest of this paper is organized as follows. Section II gives a brief description on related work and the contributions of this paper. The related work introduces tracking approaches for both general purpose and pedestrian tracking, as well as recent development in grouping-integrated pedestrian tracking. Section III describes our proposed approach in detail. Section IV shows the experimental results and provides more detailed analysis and discussion of the results. Finally this paper is concluded in Section V.

II. Related Work and Contributions

A. Related Work

Most of the state-of-the-art tracking approaches can be categorized as model-free methods, which are based on sophisticated classification approaches, for example, the online AdaBoost [15], multiple instance learning [16], and struck (structured output tracking with kernels) [17] trackers. Based on the observation that a single tracker may not always work well, a symbiotic tracker ensemble is proposed in [18]. This fusion scheme is the first attempt on bypassing the necessity of knowing all the details about an individual tracker, and achieves optimal tracking performance. One of the most important advantages of tracking approaches in this category is that they do not require a well-defined region-of-interest, but only a patch that is user defined or automatically detected for initialization. At the beginning of tracking, an online classifier is trained based on the features extracted from the initial patch. Then, for later frames, this classifier is adopted using a sliding window technique and those maximal outputs on the frames are used to locate the targeted object. In most cases, the models for tracked objects are built on appearance and shape features [2], [14]–[17]. The online design for the classifier ensures that it has the ability to update the model according to the most recent tracking results, because the model may change under different situations as time passes (e.g., illumination or pose change and so on). Meanwhile, additional strategies are adopted in these approaches to increase the tracking performance, especially for some particular situations. One of the most straightforward and useful enhancements is to use group information when tracking multiple targets together. For example, the SPOT [14] utilizes the structure information for all the targets in a similar way as in [19] to help improve performance.

For pedestrian tracking, however, methods specifically designed for pedestrians only can be used to strengthen the original tracking-by-classification approaches. For example, pedestrians can be distinguished from general objects based on the shapes of pedestrians by detection-based approaches [19], [20], and the same pedestrian can be reidentified under different environments and camera views according to his/her characteristics [21]. Since pedestrians are usually walking on the same ground plane, one common setting used in surveillance systems is to deploy multiple cameras with overlapping FOVs, so that information from multiple views can be gathered and combined [6], [7], [13]. For each pedestrian in this scenario, the images captured by different cameras may appear completely different, because the perspectives of cameras differ from each other. As a result, some difficulties in single-camera tracking systems, such as occlusions, can be overcome more easily by gathering and combining information from multiple cameras. Thus, the probability that each pedestrian can be confidently tracked in at least one camera becomes significantly higher.

From a different point of view, the structures of crowds have received more attention, since one of the most important characteristics of human beings is social interaction with one another. In other words, group formation, when people are walking together, is one of the most natural and widely observed social behaviors. This social behavior has been studied from different perspectives. For example, the group structure has been analyzed in detail [22], and the application of groups in crowd simulation has demonstrated its own value [23]. In the field of computer vision, many researchers have also shown that the spatial and temporal relationships among pedestrians are useful in enhancing the performance of a tracking system, especially when the appearance and shape features are not reliable. For example, the data association between tracklets can be greatly improved by different grouping strategies, and thus, boosting the tracking performance [5]. The pedestrian tracking in a nonoverlapping video network, in which different camera views are captured under different conditions, can also be improved by using group information [24]. Even for general objects (not pedestrians) that do not necessarily have the characteristics of sociability, the integration of group structure can still help in improving the tracking performance [14]. Although related to the method in [14], our approach is different. In particular, in [14], only 2D information is used for constructing spatial relationship

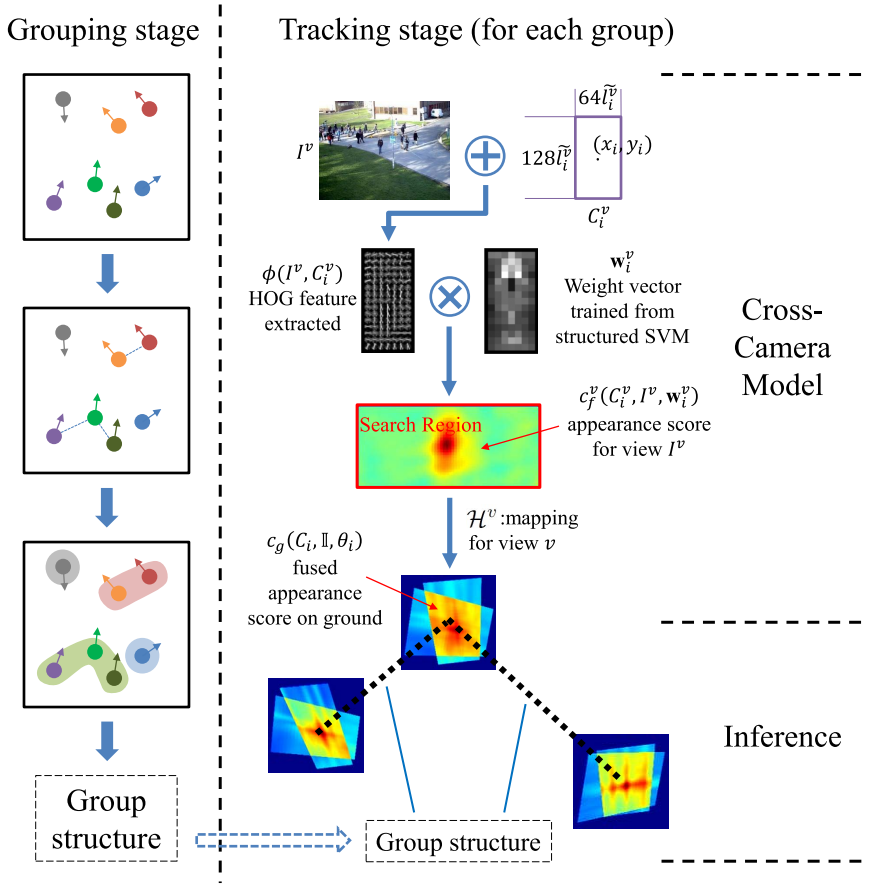


Fig. 1. Framework of the proposed method. At each time step, the groups in the crowd are first determined by the grouping approach, and the structure for each group is obtained (grouping stage), and then, the tracking is performed on each group using its structural information (tracking stage). The appearance scores are normalized to [0, 1] and shown using different colors (blue for 0 and red for 1).

among objects, while in our approach, we extend this to 3D by using a cross-camera model and project the object locations onto the ground plane for grouping purpose.

B. Contributions of This Paper

As compared with the state-of-the-art tracking systems, the contributions of this paper are as follows.

1) The multicamera pedestrian tracking system uses a two-stage tracking on the ground plane: a grouping stage and a tracking stage. In the grouping stage, the groups within the crowd are formed, and the structure for each group is represented using a minimum spanning tree. The tracking stage is extended from the SPOT [14]. For each group, its structure is preserved, and the locations for the whole group as well as for each pedestrian in the group are determined at the same time.
2) A new cross-camera model is proposed for the tracking system. For each pedestrian, one appearance model is maintained for each view, and the information from all camera views is fused together on the ground plane before the tracking stage. After the locations for the pedestrians are determined, the updated locations are then projected back to all views to update the appearance models, if necessary.
3) The tracking system is tested on challenge data. As compared with our preliminary work [25], in this paper, the differences with and without grouping stage are analyzed in detail. Furthermore, the performance under different crowd densities is discussed.

III. Technical Approach

Fig. 1 shows the framework for the entire tracking system, which can be mainly divided into two stages for each time step: 1) the grouping stage that forms the groups for all the pedestrians in a frame based on their location and velocity information from the previous time step, and computes the structure for each group and 2) the tracking stage that locates each group and all its pedestrians at the same time, and update the cross-camera model for each pedestrian who has a high confidence. The group-based tracking is derived from SPOT [14]. In this section, we will first revisit the original SPOT approach, and then we describe our improvements and contributions in detail. Table I summarizes the important notations used in this section. Note that the subscript and the superscript (e.g., i and v indicating pedestrian index and camera index in B_i^v) are explicitly expressed only when necessary for more concise presentation.

A. Structure Preserving Object Tracking

The SPOT approach is designed for single-camera generic object tracking when there are multiple objects (or patches) in the scene. It maintains the structural information

TABLE I
SUMMARY OF IMPORTANT NOTATIONS

Notation	Description
G	The group of objects (pedestrians)
$\mathbf{x}$	Object (pedestrian) location on a frame
B	Bounding box for an object (pedestrian)
C	Configuration, defined as a set of bounding boxes
I	The current frame
$\mathbf{w}$	Weight vector corresponding to a bounding box
θ	Parameter set for weight vectors
Θ	Parameter set for edges
$\vec{p}$	Pedestrian location on the ground plane
$\vec{v}$	Pedestrian velocity on the ground plane
r	Pedestrian radius on the ground plane.
$\mathbb{C}$	The set of all configurations
$\mathbb{I}$	The set of frames from all camera views

(spatial relationship) on all objects and this information is further used in determining their new locations. The approach can be generally divided into three steps: 1) compute a confidence map for each object; 2) track objects together using their confidence maps; and 3) for each object, update its corresponding classifiers using structured SVM. The feature used in the approach is the histogram of oriented gradients (HOG) feature with contrast normalization. In the following, the model, inference, and learning of the SPOT approach will be introduced. Interested readers are referred to [14] for more details on SPOT approach.

1) Model: Since it is designed for tracking general objects using a single camera, the representation for the bounding box for each object needs four parameters $B_i = \{\mathbf{x}_i, w_i, h_i\}$, where $\mathbf{x}_i = (x_i, y_i)$ indicates the location within a frame and (w_i, h_i) is the size information (width and height). In this approach, all the objects that need to be tracked are considered as in a single group G, i.e., for each object $i, i \in G$. Then, the set of bounding boxes for all objects is defined as a configuration, $C = \{B_i | i \in G\}$. The structure of this group is represented by a spanning tree, whose edges are represented by an indicator function $\mathcal{E}(\cdot, \cdot)$, where $\mathcal{E}(i, j) = 1$ means that the edge between objects i and j is included in this tree. Given a frame I, the feature extraction for each bounding box B_i is represented as $\phi(I, B_i)$. The output for the function $\phi(\cdot, \cdot)$ is a concatenated feature vector, i.e., HOG feature in this case.

For each object i, there is a corresponding weight vector $\mathbf{w}_i$, which is initialized using an SVM classifier based on the features extracted from the first frame when this object appears. Therefore, at each time step, the appearance score for each object i that measures the similarity between an observed image patch and the current object model is computed as $\mathbf{w}_i^T \phi(I, B_i)$.

The integration of structural information among all the objects is represented by adding the edge constraints between objects. Thus, the complete score of a configuration C at any frame is given by

$$S(C, I, \theta) = \sum_{i \in G} \mathbf{w}_i^T \phi(I, B_i) - \lambda_{ij} \sum_{\mathcal{E}(i,j)=1} \|(\mathbf{x}_i - \mathbf{x}_j) - e_{ij}\|^2 \tag{1}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the current locations of objects i and j, e_{ij} is a vector indicating the previous spatial relationship between them, λ represents the penalty for the spatial deformation, and θ is the set of all parameters, $\theta = \{\mathbf{w}_i | i \in G\} \cup \{e_{ij} | i, j \in G, \mathcal{E}(i, j) = 1\}$.

2) Inference: At each time step, the optimal configuration C^* that maximizes (1) needs to be obtained given the parameter set θ. In particular, this optimization can be performed in linear time using a combination of dynamic programming and min-convolution for a tree-structured graph [14]. In the tree-structured graph, each edge is formed by a node and its parent. Thus, a message-passing procedure can be used to pass information from object i to its parent node j using

$$R_{ij}(\mathbf{x}_i) = \mathbf{w}_i^T \phi(I, B_i) + \sum_{\forall k \neq j : \mathcal{E}(k,i)=1} \mu_{k \to i}(\mathbf{x}_i) \tag{2}$$

$$\mu_{i \to j}(\mathbf{x}_j) = \max_{\mathbf{x}_i'} \left(R_{ij}(\mathbf{x}_i') - \lambda \left\| (\mathbf{x}_j - \mathbf{x}_i') - e_{ij} \right\|^2 \right). \tag{3}$$

For each object i, its message that is sent to object j is determined by three facts: 1) the appearance score at bounding box B_i; 2) all the incoming messages from its child nodes; and 3) the regulation from the spatial relationship between i and j. For a tree-structured graph, this message passing starts from the root node, and the optimization finishes after a full forward–backward pass along the tree.

3) Learning: After the optimal configuration C^* is obtained, the appearance model $\mathbf{w}_i$ for each object may need to be updated as well. In [14], the learning part is conducted based on a structured SVM. The optimal configuration is considered as a true positive example in the learning process. A margin function $\Delta(C, C^*)$ is defined for the structured SVM, based on the overlapping rate between the two configurations, that is

$$\Delta(C, C^*) = \sum_{i \in G} \left(1 - \frac{B_i \cap B_i^*}{B_i \cup B_i^*} \right). \tag{4}$$

The output of the function is limited to $[0, |G|]$, where 0 can be reached if and only if $C = C^*$. Then, the loss of the structured SVM can be defined as

$$\mathcal{L}(\theta, I, C^*) = \max_C (S(C, I, \theta) - S(C^*, I, \theta) + \Delta(C, C^*)). \tag{5}$$

Since (5) only contains a set of affine functions without quadratic terms, it is a convex function with respect to the parameter set θ. Therefore, a gradient-based learning schema can be employed to solve this problem. Let us define the configuration $\bar{C}$ that provides the most negative result with respect to the optimal configuration C^* as

$$\bar{C} = \arg\max_C (S(C, I, \theta) + \Delta(C, C^*)). \tag{6}$$

Then, the general form for the gradient (with respect to θ) of the loss function in (5) can be obtained as

$$\nabla_\theta \mathcal{L}(\theta, I, C^*) = \nabla_\theta S(\bar{C}, I, \theta) - \nabla_\theta S(C^*, I, \theta). \tag{7}$$

However, sometimes, this gradient does not behave very well. To elaborate, we want to distinguish each object in the scene not only from the background but also from other objects, whereas because of the penalty, the most negative

configuration $\bar{C}$ intends to avoid using the same set of bounding boxes in C^* with a different structure. So in most cases, for each object, its most negative configuration cannot be bounding boxes for other objects. Thus, $\bar{C}$ cannot provide enough information to update object models, so that they can distinguish themselves from each other. To avoid this, a modified gradient equation is used, in which the computation of configuration score does not contain the structure penalty. The modified search direction $\mathbf{p}$ is defined as

$$\mathbf{p} = \nabla_\theta \tilde{S}(\bar{C}', I, \theta) - \nabla_\theta S(C^*, I, \theta) \tag{8}$$

where the modified computation for the most negative configuration $\bar{C}'$ is

$$\bar{C}' = \arg\max_C (\tilde{S}(C, I, \theta) + \Delta(C, C^*)) \tag{9}$$

and the confidence score is only based on the appearance score

$$\tilde{S}(C, I, \theta) = \sum_{i \in G} \mathbf{w}_i^T \phi(I, B_i). \tag{10}$$

It is easy to show that this search direction $\mathbf{p}$ and the true gradient are on the same direction, since $\mathbf{p} \cdot \nabla_\theta \mathcal{L}(\theta, I, C^*) > 0$. In this case, the learning process will still converge [14]. The configuration $\bar{C}'$ can be obtained efficiently using the same inference approach, as described in Section III-A2, except for the regulation from the spatial relationship.

The search direction is then used to update the parameter set, with a controlling parameter K, such that

$$\theta \leftarrow \theta - \frac{\mathcal{L}(\theta, I, C^*)}{\|\mathbf{p}\|^2 + \frac{1}{2K}} \mathbf{p}. \tag{11}$$

The parameter set for object i is only updated when the confidence at the patch B_i is larger than a threshold T_c.

B. Group Structure Preserving Pedestrian Tracking

Although the SPOT tracking approach [14] is able to utilize the structural information of groups, it needs several necessary extensions before it is suitable to be used in our multicamera pedestrian tracking system. The main extensions are: 1) the incorporation of the grouping stage before the tracking stage and 2) the proposal of the cross-camera model for pedestrians. By adding the grouping stage, the crowd of pedestrians is divided into several groups rather than a single group in the original approach [14], and the tracking is done for each group separately. The cross-camera model maps and fuses each pedestrian's information from all different cameras onto the ground plane; thus, it builds a confidence score map for each pedestrian on the ground plane. This works as the view-based confidence map in the original SPOT tracker [14]. In the following, we describe these extensions in detail.

1) Grouping Stage: A major limitation of the SPOT tracking approach [14] is that it does not have an automatic grouping step. The groups of objects have to be manually assigned and their structures need to be precomputed before tracking. In addition, the group information is fixed during the entire tracking process. Therefore, a grouping stage is added in our pedestrian tracking system to make the system applicable to the complex real-world scenarios.

We have adopted a state-of-the-art pedestrian tracking method [5] for our grouping stage. In [5], the pedestrians are grouped based on their current status, i.e., their locations as well as velocities. Note that our pedestrian tracking system works in a video network that consists of multiple cameras with overlapping FOVs and we use homography to map image planes to the ground plane. Therefore, using the location and velocity information on the ground plane for all pedestrians in our tracking system becomes a better choice than using the information on image planes.

For each pedestrian i, his/her status on the ground plane is represented as $(x_i^g, y_i^g, u_i^g, v_i^g)$, where $\vec{p}_i = (x_i^g, y_i^g)$ indicates the location and $\vec{v}_i = (u_i^g, v_i^g)$ indicates the velocity. A pairwise grouping score is computed between every two pedestrians based on their spatial relationship, according to their locations and velocities, that is

$$S_{ij}^g = D_{ij}^g \cdot V_{ij}^g \tag{12}$$

where D^g and V^g are the scores computed based on the location and the velocity relationship between pedestrians i and j, respectively. The two scores are calculated using

$$D_{ij}^g = 1 - \frac{2}{\pi} \arctan(\text{dist}(\vec{p}_i, \vec{p}_j)) \tag{13}$$

$$V_{ij}^g = \frac{1}{\exp(\text{vel}(\vec{v}_i, \vec{v}_j)) + 1} \tag{14}$$

where $\text{dist}(\vec{p}_i, \vec{p}_j)$ is a relative distance between the two pedestrians and $\text{vel}(\vec{v}_i, \vec{v}_j)$ is the relative velocity between them. These two items are computed by

$$\text{dist}(\vec{p}_i, \vec{p}_j) = \max\left(0, \frac{\|\vec{p}_i - \vec{p}_j\|^2}{r_i + r_j} - 1\right) \tag{15}$$

$$\text{vel}(\vec{v}_i, \vec{v}_j) = 2\frac{\|\vec{v}_i - \vec{v}_j\|^2}{r_i + r_j} \tag{16}$$

where r_i and r_j are the radius for pedestrian i and j, respectively. For simplicity, we assume that all the pedestrians have the same radius (size) in our system, that is, $r_i = r_j = r$. The operations in the definition of $\text{dist}(\cdot, \cdot)$ ensure that the smallest distance between any two pedestrians is the sum of their radii $(2r)$. The range of D^g and V^g is $(0, 1]$.

From the above equations, we can observe a significant difference between our grouping strategy and the strategy in [5]. Since the grouping strategy in [5] aims to group tracklets, which consist of information from multiple frames, the grouping score for each frame is computed and then averaged over all frames to obtain the grouping score between two tracklets. However, in our tracking system, we adopt a grouping strategy that supports an instant decision, which means that at each time step, the groups are determined based only on the information from the previous time step. This strategy is more robust against the change in the number of groups.

With these pairwise grouping scores, the labels that indicate whether two pedestrians are in the same group or not can be obtained using a threshold T_g. When two pedestrians i and j have a grouping score $S_{ij}^g \geq T_g$, the indicator function $\mathcal{E}_{(i,j)}$ is set to 1, otherwise 0. Then, the groups of all pedestrians are

extended from these pairwise connectivities. If two pedestrians i and j are connected [$\mathcal{E}_{(i,j)} = 1$], and i and k are connected as well [$\mathcal{E}_{(i,k)} = 1$], then we label i, j, and k as belonging to the same group. This step iterates until the groups are converged. As a result, given the set of all pedestrians $\mathbb{P} = \{i\}$, a group of pedestrians is defined as a set of pedestrians $G_i = \{i_1, i_2, \ldots, i_n | 1 \leq i_k \leq |\mathbb{P}|\}$ that forms a connected graph. In addition, any two groups G_i and G_j in the set of groups $\mathbb{G} = \{G_i\}$ are nonoverlapping, $G_i \cap G_j = \emptyset (\forall i, j)$. The structure for each group G is computed over its edge set $E = \{e_{ij} | i, j \in G, \mathcal{E}_{(i,j)} = 1\}$, and the weight of an edge is defined as the Euclidean distance between two pedestrians. The SPOT tracking [14] is then conducted for these groups. Note that this grouping stage is performed at each time step before the tracking stage, the groups, and their structural information may be different from frame to frame.

2) Cross-Camera Model: The object model in [14] also needs to be extended to utilize and update information from all camera views. Similar to the original model, we define a bounding box $B_i^v = (x_i^v, y_i^v, w_i^v, h_i^v)$ for each pedestrian i in each camera view $v (v \in \mathbb{V})$, where $\mathbf{x}_i^v = (x_i^v, y_i^v)$ is the frame location and (w_i^v, h_i^v) is the size information (width and height).

For generic object tracking, the size information of a bounding box is usually unpredictable, because the positions of the camera and the targeted object, as well as the actual 3D shape of the object, are arbitrary. However, in pedestrian tracking systems, the targets are usually walking pedestrians, with a similar shape from almost all perspectives (except to those images that were captured in a bird view). Therefore, we use a scaling technique to represent the size of the bounding box for a single pedestrian. In particular, we set a standard size for all pedestrian bounding boxes as (w, h) and the scale of an arbitrary pedestrian bounding box is computed as $l_i^v = h_i^v / h$. That is, the size of the pedestrian's bounding box can be represented as $(w \cdot l_i^v, h \cdot l_i^v)$, using only one scaling variable l_i^v. As a result, the status for a bounding box B_i^v in our system can be redefined as $B_i^v = (x_i^v, y_i^v, l_i^v)$.

Accordingly, for each pedestrian i, a configuration is defined as all corresponding bounding boxes for all camera views $C_i = \{B_i^v | v \in \mathbb{V}\}$, and a configuration for all the pedestrians in a group $G \in \mathbb{G}$ is then defined as $\mathbb{C} = \{C_i | i \in G\}$.

Another difference between this cross-camera model and the original model used in [14] is the feature extraction. The original model uses the HOG feature with contrast normalization. However, this feature descriptor by itself is not sufficient in a pedestrian tracking system, since it may fail to distinguish one pedestrian from another. Therefore, in the current approach, we also incorporate color feature in addition to the original HOG feature. Given a bounding box B_i^v and a frame I^v, we first resize the bounding box and the frame according to the scale l_i^v, so that the bounding box is resized to the standard size (w, h). The HOG feature and the color feature are then extracted on the resized frame. We use $\Phi(I^v, B_i^v)$ to indicate this new feature extraction process. The updated feature extraction concatenates the color features and the HOG features.

For a pedestrian i in camera view v, the appearance score on this camera view can be calculated as

$$c_f^v\left(B_i^v, I^v, \mathbf{w}_i^v\right) = \mathbf{w}_i^{v\,T} \cdot \Phi\left(I^v, B_i^v\right) \tag{17}$$

where $\mathbf{w}_i^v$ is the trained weight vector on the features extracted using structured SVM. Then, the fused appearance scores on the ground plane can be obtained by

$$c_g(C_i, \mathbb{I}, \theta_i) = \frac{1}{\|\mathbb{V}\|} \sum_{v \in \mathbb{V}} \mathcal{H}^v\left(c_f^v\left(B_i^v, I^v, \mathbf{w}_i^v\right)\right) \tag{18}$$

where $\mathbb{I}$ is defined as the set of all frames across all views $\mathbb{I} = \{I^v | v \in \mathbb{V}\}$ and $\theta_i = \{\mathbf{w}_i^v | v \in \mathbb{V}\}$. $\mathcal{H}^v(\cdot)$ is a projection function that transforms the appearance scores at the frame coordinates of camera view v to the scores at the corresponding coordinates on the ground plane. Since most of the cameras used in pedestrian tracking systems have perspective views, we use homography to project the feet locations of all pedestrians.

Therefore, for a configuration $\mathbb{C}$, its complete score can be calculated using the following equation corresponding to (1):

$$\begin{aligned} S_g(\mathbb{C}, \mathbb{I}, \Theta) = &\sum_{i \in G} c_g(C_i, \mathbb{I}, \theta_i) \\ &- \lambda_{ij} \sum_{\mathcal{E}(i,j)=1} \|(\vec{p}_i - \vec{p}_j) - e_{ij}\|^2 \end{aligned} \tag{19}$$

where e_{ij} is the edge vector indicating the location difference for pedestrians i and j on the ground plane from the previous time step. Θ is the set of all parameters, $\Theta = \{\theta_i | i \in G\} \cup \{e_{ij} | i, j \in G, \mathcal{E}(i, j) = 1\}$.

Using the inference similar to Section III-A2, the optimal configuration $\mathbb{C}^*$ for all the pedestrians in a group can be obtained. The only difference is that all the calculations in the inference are done on the ground plane. In other words, the fused appearance score $c_g(C_i, \mathbb{I}, \theta_i)$ and the location $\vec{p}_i$ on the ground plane are used instead.

The model updating or model learning process is also different from the original work [14], since the models for all camera views need to be updated together. The margin function for the structured SVM is extended to all camera views, that is

$$\Delta(\mathbb{C}, \mathbb{C}^*) = \sum_{v \in \mathbb{V}} \sum_{i \in G} \left(1 - \frac{B_i^v \cap B_i^{v*}}{B_i^v \cup B_i^{v*}}\right). \tag{20}$$

The loss function is defined according to the complete scores on the ground plane and the extended margin function

$$\mathcal{L}(\Theta, \mathbb{I}, \mathbb{C}^*) = \max_{\mathbb{C}} (S_g(\mathbb{C}, \mathbb{I}, \Theta) - S_g(\mathbb{C}^*, \mathbb{I}, \Theta) + \Delta(\mathbb{C}, \mathbb{C}^*)). \tag{21}$$

It may seem that this extended loss function is more complex than the original one in (5). However, the calculations of this loss function are still limited to a set of affine functions, without any quadratic terms. Therefore, it is still a convex function with respect to the parameter set Θ, which is the same as in the original work [14].

With this loss function, the learning procedure is essentially the same as the procedure described in Section III-A3, except

that all the scores used are the fused scores on the ground plane. The search direction is defined as

$$\mathbf{p}_g = \nabla_\Theta \tilde{S}_g(\bar{\mathbb{C}}', \mathbb{I}, \Theta) - \nabla_\Theta S_g(\mathbb{C}^*, \mathbb{I}, \Theta) \tag{22}$$

where $\bar{\mathbb{C}}'$ and $\tilde{S}_g$ are computed as

$$\bar{\mathbb{C}}' = \arg\max_{\mathbb{C}} (\tilde{S}_g(\mathbb{C}, \mathbb{I}, \Theta) + \Delta(\mathbb{C}, \mathbb{C}^*)) \tag{23}$$

$$\tilde{S}_g(\mathbb{C}, \mathbb{I}, \Theta) = \frac{1}{\|\mathbb{V}\|} \sum_{v \in \mathbb{V}} \sum_{i \in G} \mathbf{w}_i^{v\,T} \phi\left(I^v, B_i^v\right). \tag{24}$$

Since the weight vectors $\mathbf{w}_i^v$ are concatenated in the parameter set Θ, they can also be updated using (11), with the global information for $\mathcal{L}(\Theta, \mathbb{I}, \mathbb{C}^*)$ and $\mathbf{p}_g$ provided

$$\Theta \leftarrow \Theta - \frac{\mathcal{L}(\Theta, \mathbb{I}, \mathbb{C}^*)}{\|\mathbf{p}_g\|^2 + \frac{1}{2K}} \mathbf{p}_g. \tag{25}$$

3) Complexity Analysis: The original SPOT tracker [14] needs the confidence map for each object [$O(n)$ where n is the number of objects] and then accomplishes the group tracking based on inference [$O(n)$]. So the complexity for the complete system is still $O(n)$.

Compared with the original SPOT tracker, the proposed approach has an additional grouping stage and extends the object model to a cross-camera model, which uses information from all camera views. In the grouping stage, the computation of pairwise grouping score S_{ij}^g and the determination of the group structure require a complexity of $O(n^2)$ (n is the number of pedestrians). In the cross-camera model, the computation of the confidence map for each view for each pedestrian is the same as in the original SPOT tracker [14], which is $O(n)$, so the calculation of the confidence score on the ground plane is $O(|\mathbb{V}|n)$ where $|\mathbb{V}|$ is the number of views. Therefore, in the proposed approach, the overall complexity becomes $O(n^2 + |\mathbb{V}|n)$. Note that the computational cost for the confidence score map on the ground plane is necessary as we need to gather information for all pedestrians from all camera views.

IV. Experiments

In order to evaluate the proposed group structure preserving pedestrian tracking approach, experiments are conducted on several challenging sequences from publicly available data sets. In addition, to provide further evidence of the effectiveness of the proposed approach, more detailed analyses are provided. We describe the experimental results as well as their corresponding discussions in this section.

A. Experimental Settings

1) Data: We use two sequences for the experiments, both from the PETS 2009 data set. The first sequence is the S2.L1 (795 frames), containing low density crowd in the scenario. Another sequence is the S2.L2 (435 frames) that has medium density crowd. For the first sequence, Views 1, 5, and 7 are used for tracking. For the second sequence, frames from Views 1 and 2 are used in our experiments. For each view, we manually annotate the ground truth for every five frames; the ground truth in between frames is obtained using linear interpolation. On the ground plane, the ground truth is computed using principal-axis-based correspondence [26] for all camera views.

2) Implementation Details: In the original SPOT tracking approach [14], two tree models are tested: a minimum spanning tree model and a star model. In our implementation, we only use the minimum spanning tree model. The reasons are twofold: 1) our grouping output provides an initial edge set for each group, which means that two pedestrians in the same group may not be directly related and 2) the original work [14] shows that the minimum spanning tree model has a better performance than the star model.

Moreover, since the inference for each group is performed on the ground plane, the computation is rather time-consuming when the ground plane is divided into a dense grid (details in Section IV-B). Therefore, the appearance score for each pedestrian for each view is only computed in a small search region around its previous location. As a result, the fused appearance score on the ground plane can also be limited to a small region. The size of the search region for each pedestrian on each camera view is set to 320×240 with the bounding box resized and centered at the previous frame location of this particular pedestrian. This choice of size is able to significantly reduce the computation time, while it is not too small for recovery from potential drifting. The location for pedestrians on the ground plane is obtained using the principal-axis-based computation [26].

In addition, for each bounding box, not all possible scales are used in appearance score calculation, since the scale change for a walking pedestrian from frame to frame is relatively small. In the experiment, we use three scale levels: 0.95, 1, and 1.05. The appearance score at each position is the largest one among the scores computed under these three levels. For each pedestrian, each tracker is initialized using manual annotation to avoid unnecessary errors. A tracker is considered as inactive if its score cannot exceed the threshold for more than ten consecutive frames. This means that the parameter set Θ of the tracker is not updated.

3) Parameter Settings: The standard size of each pedestrian patch is set to 64×128, which is commonly used in pedestrian detection. The ground plane is set to a grid with a size of 700×700. The size of each cell in the grid is $\sim$6 cm $\times$ 6 cm, which is determined empirically. For each view v, we manually label four corresponding points on the image plane and the ground plane to estimate the projection functions $\mathcal{H}^v(\cdot)$. The estimated radius of pedestrians r is set to five cells (about 30 cm).

For the parameters, we follow the settings from the original work [14]. That is, λ_{ij} in (19) is set to 0.001, the confidence threshold T_c is set to 0.4, and the control parameter K in the model learning (11) is set to 1. We have empirically found that these parameter settings provide satisfactory results. The training of the initial weight vector $\mathbf{w}_i^v$ uses libSVM [27] with precomputed kernel matrix and default parameters. The grouping threshold T_g is set to 0.2 in all experiments.

B. Results

In this section, the results for the two complete sequences S2.L1 and S2.L2 from PETS 2009 are provided first, followed by the detailed analysis.

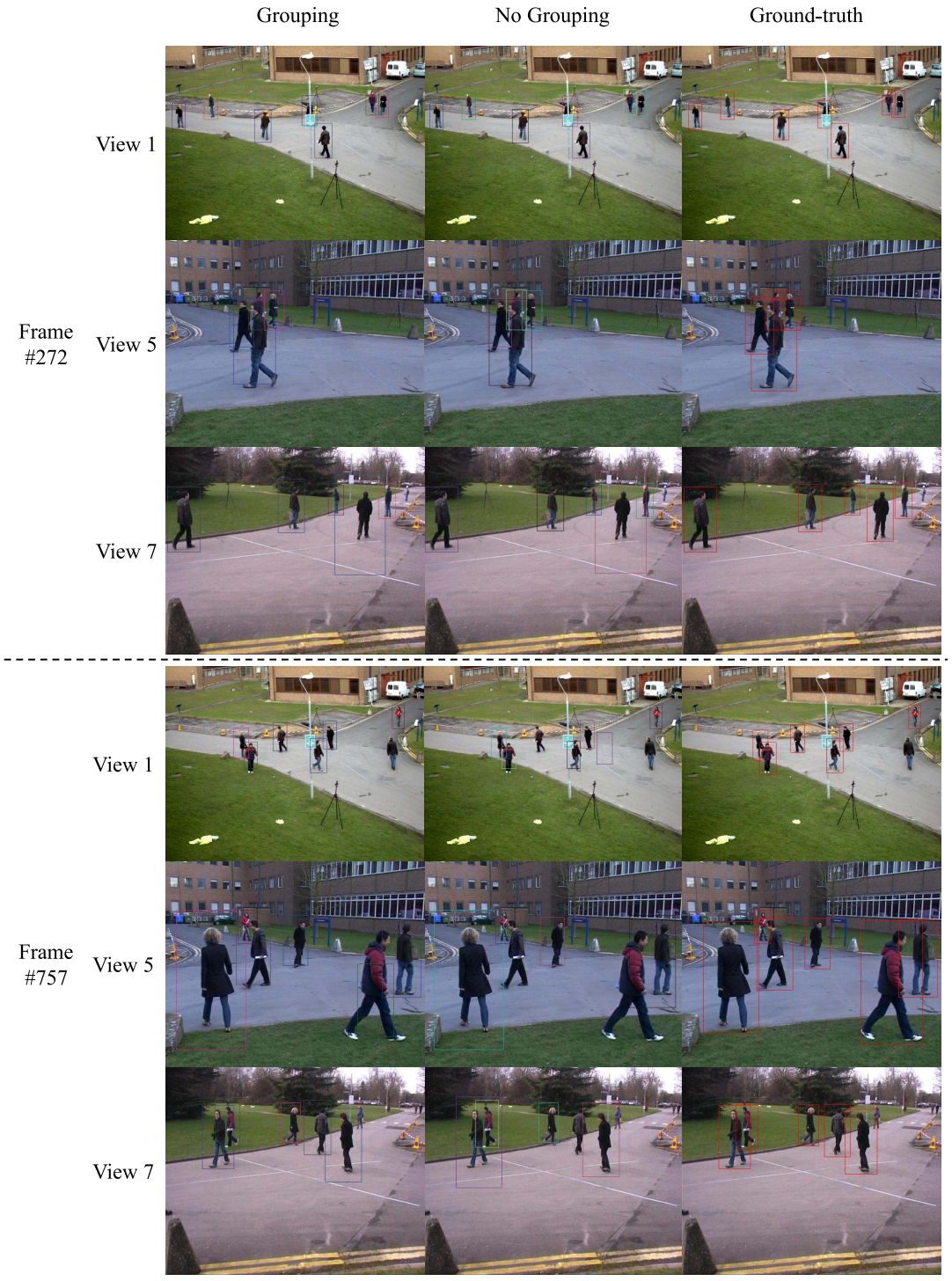


Fig. 2. Sample frames for sequence S2.L1. Two frames (#272 and #757) with Views 1, 5, and 7 are shown. The left two columns show the tracking results with and without grouping, respectively. The right column contains the corresponding ground truth. The bounding boxes with the same color in different views are for the same pedestrian.

1) Complete Sequences: Fig. 2 shows the results for frames #272 and #757 from the sequence S2.L1 Views 1, 5, and 7. Fig. 3 shows frames #065 and #420 from the sequence S2.L2 Views 1 and 2. The first column shows the tracking results with grouping; the frames in the second column are the results obtained when tracking pedestrians without grouping, and the third column provides the ground truth.

We use multiobject tracking precision (MOTP) and multiobject tracking accuracy (MOTA) [29] to quantitatively evaluate the tracking performance. The evaluation is conducted on both the camera views and the ground plane.

Since the information for each pedestrian is represented as a bounding box in each camera view, the accuracy of the tracker is computed based on the overlapping ratio between the tracked bounding box and the ground truth. The overlapping

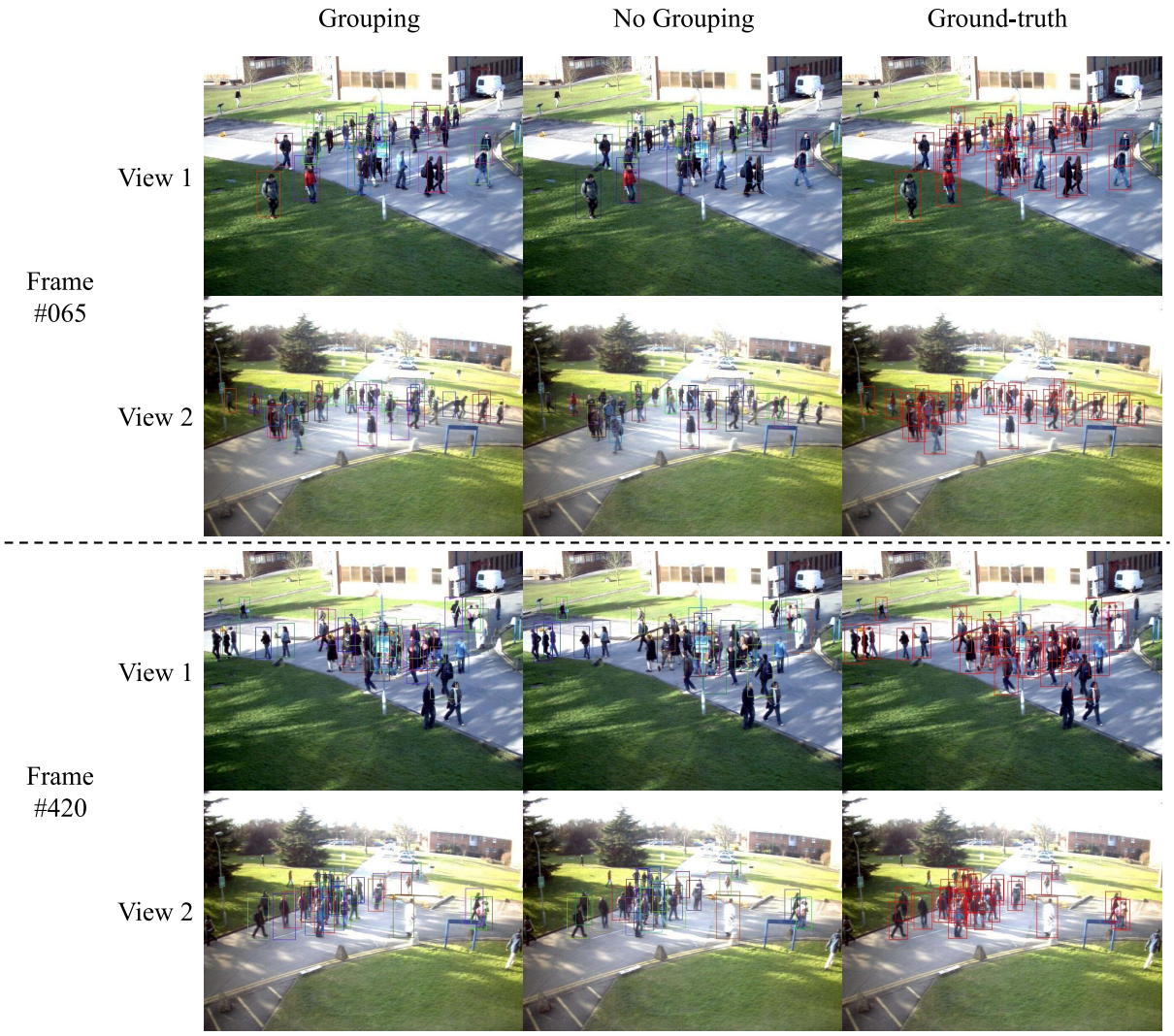


Fig. 3. Sample frames for sequence S2.L2. Two frames (#065 and #420) with Views 1 and 2 are shown. The left two columns show the tracking results with and without grouping, respectively. The right column contains the corresponding groundtruth. The bounding boxes with the same color in different views are for the same pedestrian. The total number of pedestrians in this sequence varies between 10 and 36.

ratio of two bounding boxes is defined as the area of their intersection divided by the area of their union, which falls in the range of [0, 1]. The tracking is considered to be accurate when the overlapping ratio is greater than 0.5.

On the ground plane, however, the information for each pedestrian is maintained using point-based location. Thus, we use a Euclidean distance-based calculation for accuracy computation

$$d(\text{tr}, \text{gt}) = \max\left(0, 1 - \frac{\|p_{\text{tr}} - p_{\text{gt}}\|^2}{4r}\right). \tag{26}$$

When $d(\text{tr}, \text{gt}) > 0.5$, the tracked location is considered to be accurate. That is, the distance between the tracked location (p_{tr}) and the ground truth (p_{gt}) is smaller than the diameter of a pedestrian.

Table II shows the MOTP and MOTA evaluation results for the two sequences. The results reveal that the structural information for groups improves the performance of the tracking system under different crowd densities. In addition, the improvement is more obvious under medium density crowd scenarios (S2.L2) compared with low-density crowd scenarios (S2.L1). This is reasonable, since in general, more potential groups exist when crowd density gets higher, which means that the structural information of groups is more useful.

Compared with our previous tracking system [28] that is integrated with a crowd simulator, the evaluation results on the ground plane are better for both the sequences. For sequence S2.L2, the results on camera views from the current system are similar to [28]; but for sequence S2.L1, the evaluation results on Views 5 and 7 are not as good as those in [28]. The main reason for this is that the sizes of the bounding boxes in the current system cannot be drastically changed (shown in Fig. 2, Views 5 and View 7), while in [28], the sizes are determined by pedestrian detections. Therefore, a bounding box, which may still be centered at the correct position, may not be evaluated as accurate, since its overlap with the ground truth is small. However, this does not have significant

TABLE II
MOTP AND MOTA EVALUATION FOR TWO SEQUENCES

Sequence	Evaluation Category	No Grouping		Grouping		Method in [28]	
		MOTP	MOTA	MOTP	MOTA	MOTP	MOTA
S2.L1	View 1	77.14%	85.29%	**78.99%**	**89.20%**	76.49%	88.86%
	View 5	72.27%	78.79%	**72.98%**	82.69%	72.89%	**85.82%**
	View 7	73.89%	77.62%	**75.41%**	80.86%	74.98%	**85.77%**
	Ground plane	**81.98%**	82.60%	81.25%	**87.62%**	80.46%	86.36%
S2.L2	View 1	**73.11%**	59.83%	72.71%	67.09%	71.53%	**67.56%**
	View 2	71.20%	57.58%	**72.16%**	**64.49%**	71.39%	63.79%
	Ground plane	80.12%	29.67%	**83.83%**	**43.03%**	78.43%	37.77%

TABLE III
MOTP AND MOTA EVALUATION FOR ALL THREE SEGMENTS FROM PETS 2009 S2.L2

Frames	Evaluation Category	No Grouping		Grouping	
		MOTP	MOTA	MOTP	MOTA
#000-#019	View 1	81.12%	**74.30%**	**81.60%**	**74.30%**
	View 2	**82.92%**	78.11%	82.44%	**80.43%**
	Ground plane	**85.04%**	59.20%	84.71%	**66.83%**
#115-#134	View 1	81.58%	**71.08%**	**82.32%**	70.10%
	View 2	**83.05%**	74.02%	82.90%	**84.80%**
	Ground plane	86.49%	56.37%	**87.29%**	**63.24%**
#300-#319	View 1	82.39%	**78.80%**	**82.61%**	**78.80%**
	View 2	82.52%	69.02%	**82.82%**	**72.28%**
	Ground plane	**84.38%**	70.11%	84.26%	**75.54%**

TABLE IV
TRACKING PERFORMANCE ON SEGMENTS FROM PETS 2009 S2.L2 FOR PEDESTRIANS IN GROUPS ONLY

Evaluation Category	Tracking Individually			Tracking in Groups		
	MOTP	MOTA	Distance	MOTP	MOTA	Distance
View 1	82.13%	76.54%	167.91	**82.89%**	**79.42%**	**136.41**
View 2	81.89%	70.75%	90.36	**82.42%**	**83.42%**	**68.17**
Ground plane	84.33%	59.58%	36.11	**85.14%**	**72.71%**	**26.44**

TABLE V
AVERAGE PROCESSING TIME FOR EACH FRAME AND EACH PEDESTRIAN FOR PETS 2009 S2.L2

Pedestrian Density	$5 \sim 14$	$15 \sim 24$	$25 \sim 34$
With Grouping	98.5ms	99.1ms	99.3ms
Without Grouping	97.6ms	98.3ms	97.9ms

impact on tracking on the ground plane, since the location on the ground plane is determined by the principal axis of the bounding box. Conclusively, with the cross-camera model, the group structural information helps pedestrian tracking in a multicamera video network, especially for the pedestrian locations on the ground plane.

2) Detailed Analysis: The first experiment investigates the tracking performance under different crowd densities. We select three segments from sequence S2.L2. The first segment starts from frame #0, the second starts from frame #115, and the third starts from frame #300. Each segment has 20 frames. For the first segment, there are about 30 pedestrians in the scenario; for the second one, there are about 20 pedestrians; and for the third one, there are about 10 pedestrians. The evaluation results on these three segments are reported in Table III. We can observe that the tracking performance is improved under all three different pedestrian densities.

The second experiment aims to study the tracking performance for those pedestrians who are involved in groups only. Since sequence S2.L2 is more complicated than S2.L1, we use this sequence for this analysis. At first, the ground truth of groupings is obtained by running our grouping approach over the ground truth of pedestrian locations and velocities on the ground plane. Then, among all the groups that last more than 30 frames, we randomly select 20 groups, and for each group, we randomly select 20 continuous frames for testing. The tracking is performed on these 20 groups, and for each run, the results are evaluated only on those pedestrians who are in the corresponding group. Besides the MOTP and MOTA metrics, we use distance-based evaluation as well. The distance on each frame is computed as the pixel-based Euclidean distance between the centers of the tracked bounding box and the ground truth. The distance on the ground plane is computed as the cell-based Euclidean distance. Two different tracking strategies are investigated: tracking individually and tracking in groups. The results are reported in Table IV. The significant differences observed in this experiment suggest that the performance improvement is brought by the structural information of groups. Fig. 4 shows the sample results for tracking two pedestrians in a group using two different strategies.

C. Computational Cost

As mentioned in Section III-B3, the proposed approach has an overall complexity of $O(n^2 + |\mathbb{V}|n)$, and the computation time of confidence score is already minimal. So the only extra computation is the grouping cost. However, in reality, the computation for the confidence score is usually time consuming, so that the additional cost for the grouping may not be observable. Table V shows the average processing time for each frame and each pedestrian with/without grouping under different pedestrian densities for the data sequence PETS 2009 S2.L2. The tracking system is implemented in C++ using OpenCV 2.41 and Visual Studio 2012, and it runs on a laptop with Intel i7 2675QM 2.8 GHz and 8-GB RAM. The results show that the average processing time is almost identical under different pedestrian densities, and the difference between systems with and without grouping is very small. This means that the grouping strategy does not significantly increase the computational time, while it helps improve the tracking performance.

Fig. 4. Sample frames for tracking two pedestrians in a group using two different strategies. From the left column to the right column, four frames (#048, #053, #058, and #063) are shown. The top two rows show the tracking results with and without grouping, respectively. The last row contains the corresponding ground truth. Frames are enlarged and cropped for better illustration.

V. Conclusion

In this paper, a new approach that tracks pedestrians in a multicamera video network is proposed. A unique merit of this approach is that it preserves group structure during tracking. At each time step, the integrated grouping computes the groups of all pedestrians based on their previous locations and velocities, and calculates the structural information for each group. The extended structure preserving tracking is then used for each group for tracking pedestrians, and a new cross-camera model is used to fuse information from multiple camera views. After the inference on the ground plane has been made, the locations for all the pedestrians in the group are determined jointly, and the model for each pedestrian can be updated according to the new information. The experiments on challenging data demonstrate that the integration of group structural information can help improve the tracking performance significantly.

References

[1] B. Bhanu, C. V. Ravishankar, A. K. Roy-Chowdhury, H. Aghajan, and D. Terzopoulos, Eds., *Distributed Video Sensor Networks*. London, U.K.: Springer, 2011.

[2] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, Jun. 2008, pp. 1–8.

[3] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 3457–3464.

[4] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1820–1833, Sep. 2011.

[5] X. Chen, Z. Qin, L. An, and B. Bhanu, "An online learned elementary grouping model for multi-target tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 1242–1249.

[6] W. Du and J. Piater, "Multi-camera people tracking by collaborative particle filters and principal axis-based integration," in *Proc. 8th Asian Conf. Comput. Vis. (ACCV)*, Tokyo, Japan, 2007, pp. 365–374.

[7] R. Eshel and Y. Moses, "Tracking in a dense crowd using multiple cameras," *Int. J. Comput. Vis.*, vol. 88, no. 1, pp. 129–143, May 2010.

[8] O. Javed, K. Shafique, Z. Rasheed, and M. Shah, "Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views," *Comput. Vis. Image Understand.*, vol. 109, no. 2, pp. 146–162, 2008.

[9] X. Chen, Z. Qin, L. An, and B. Bhanu, "Multi-person tracking by online learned grouping model with non-linear motion context," *IEEE Trans. Circuits Syst. Video Technol.*, [Online]. Available: doi: 10.1109/TCSVT.2015.2511480.

[10] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 363–376, Jan. 2013.

[11] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, Jun. 2012, pp. 1815–1821.

[12] O. Ozturk, T. Yamasaki, and K. Aizawa, "Tracking of humans and estimation of body/head orientation from top-view single camera for visual focus of attention analysis," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops (ICCVW)*, Kyoto, Japan, Sep./Oct. 2009, pp. 1020–1027.

[13] Z. Jin and B. Bhanu, "Integrating crowd simulation for pedestrian tracking in a multi-camera system," in *Proc. ACM/IEEE 6th Int. Conf. Distrib. Smart Cameras (ICDSC)*, Hong Kong, Oct./Nov. 2012, pp. 1–6.

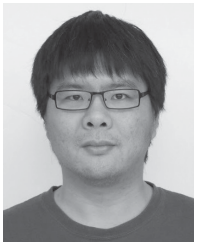
[14] L. Zhang and L. van der Maaten, "Preserving structure in model-free tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 756–769, Apr. 2014.

[15] H. Grabner and H. Bischof, "On-line boosting and vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2006, pp. 260–267.

[16] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1619–1632, Aug. 2011.

[17] S. Hare, A. Saffari, and P. H. S. Torr, "Struck: Structured output tracking with kernels," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Barcelona, Spain, Nov. 2011, pp. 263–270.

[18] Y. Gao, R. Ji, L. Zhang, and A. Hauptmann, "Symbiotic tracker ensemble toward a unified tracking framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 7, pp. 1122–1131, Jul. 2014.
[19] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. San Diego, CA, USA, Jun. 2005, pp. 886–893.
[21] L. An, M. Kafai, S. Yang, and B. Bhanu, "Person reidentification with reference descriptor," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 4, pp. 776–787, Apr. 2016.
[22] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz, "The walking behaviour of pedestrian social groups and its impact on crowd dynamics," *PLoS ONE*, vol. 5, no. 4, p. e10047, Apr. 2010.
[23] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, Jun. 2011, pp. 1345–1352.
[24] L. Feng and B. Bhanu, "Understanding dynamic social grouping behaviors of pedestrians," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 2, pp. 317–329, Mar. 2015.
[25] Z. Jin and B. Bhanu, "Multi-camera pedestrian tracking using group structure," in *Proc. ACM/IEEE Int. Conf. Distrib. Smart Cameras (ICDSC)*, Venice, Italy, 2014, Art. no. 2.
[26] W. Hu, M. Hu, X. Zhou, T. Tan, J. Luo, and S. Maybank, "Principal axis-based correspondence between multiple cameras for people tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 663–671, Apr. 2006.
[27] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm
[28] Z. Jin and B. Bhanu, "Analysis-by-synthesis: Pedestrian tracking with crowd simulation models in a multi-camera video network," *Comput. Vis. Image Understand.*, vol. 134, pp. 48–63, May 2015.
[29] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *J. Image Video Process.*, vol. 2008, pp. 1–10, Feb. 2008.



Zhixing Jin received the B.Eng. and M.Sc. degrees in computer science and technology from Tsinghua University, Beijing, China, in 2006 and 2009, respectively, and the Ph.D. degree in computer science from University of California, Riverside, CA, USA, in 2014.

His research interests include computer vision and pattern recognition.

Le An received the B.Eng. degree in telecommunications engineering from Zhejiang University, Hangzhou, China, in 2006; the M.Sc. degree in electrical engineering from Eindhoven University of Technology, Eindhoven, The Netherlands, in 2008; and the Ph.D. degree in electrical engineering from University of California, Riverside, CA, USA, in 2014.

His research interests include image processing, computer vision, pattern recognition, and machine learning.

Bir Bhanu (S'72–M'82–SM'87–F'95) received the S.M. and E.E. degrees in electrical engineering and computer science from Massachusetts Institute of Technology, Cambridge, MA, USA; the Ph.D. degree in electrical engineering from the Image Processing Institute, University of Southern California, Los Angeles, CA, USA; and the M.B.A. degree from University of California, Irvine, CA, USA.

He is currently the Distinguished Professor of Electrical and Computer Engineering, a Cooperative Professor of Computer Science and Engineering and Mechanical Engineering, and the Interim Chair of the Department of Bioengineering with University of California, Riverside, CA, USA. In addition, he is the Director of the Center for Research in Intelligent Systems, the Visualization and Intelligent Systems Laboratory, and the NSF IGERT on Video Bioinformatics. He has authored over 475 reviewed technical publications, including over 135 journal papers and 45 book chapters. He has published seven authored and three edited books. He holds 18 patents and has five pending. His research interests include computer vision, pattern recognition and data mining, machine learning, artificial intelligence, image processing, image and video database, graphics and visualization, robotics, human–computer interactions, and biological, medical, military, and intelligence applications.

Dr. Bhanu is a fellow of AAAS, AIMBE, IAPR, and SPIE. He has been the Principal Investigator of various programs for NSF, DARPA, NASA, AFOSR, ONR, ARO, and other agencies and industries in the areas of video networks, video understanding, video bioinformatics, learning and vision, image understanding, pattern recognition, target recognition, biometrics, autonomous navigation, image databases, and machine-vision applications.