Vision-based Analysis of Small Groups in Pedestrian Crowds

Weina Ge, Robert T. Collins, *Senior Member, IEEE*, and R. Barry Ruback
E-mail: gewe@ge.com, rcollins@cse.psu.edu, rbr3@psu.edu

***Abstract*—Building upon state-of-the-art algorithms for pedestrian detection and multi-object tracking, and inspired by sociological models of human collective behavior, we automatically detect small groups of individuals who are traveling together. These groups are discovered by bottom-up hierarchical clustering using a generalized, symmetric Hausdorff distance defined with respect to pairwise proximity and velocity. We validate our results quantitatively and qualitatively on videos of real-world pedestrian scenes. Where human-coded ground truth is available, we find substantial statistical agreement between our results and the human-perceived small group structure of the crowd. Results from our automated crowd analysis also reveal interesting patterns governing the shape of pedestrian groups. These discoveries complement current research in crowd dynamics, and may provide insights to improve evacuation planning and real-time situation awareness during public disturbances.**

***Index Terms*—pedestrian detection and tracking, pedestrian groups, crowd dynamics**

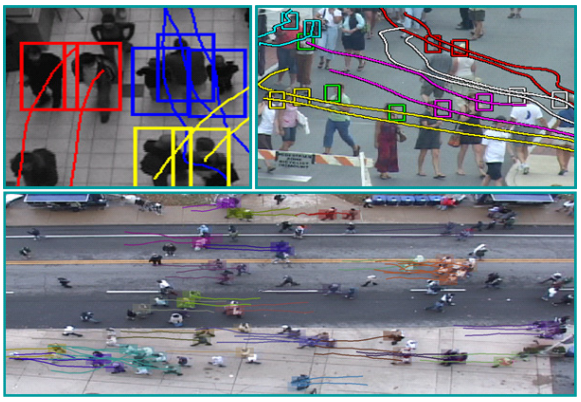

Fig. 1: Small groups are prevalent in pedestrian scenes. Our algorithm detects groups of people traveling together via hierarchical clustering on trajectories automatically extracted from video of crowds under various conditions.

I. Introduction

There has been increasing interest in using surveillance trajectory data for human behavior analysis, ranging from activity recognition based on the motion pattern of a single individual or interactions among a few (e.g. [1]), to analysis of the flow of a large crowd, for example to discover pathways or monitor for abnormal events (e.g. [2]). Less well-studied is the collective behavior of small groups of people in a crowd. In this paper we build upon state-of-the-art pedestrian detection and tracking techniques to discover small groups of people who are traveling together. Determining the group structure of a crowd provides a basis for further mid-level analysis of events involving social interactions of and between groups.

Our main contribution is a hierarchical clustering algorithm that, informed by social psychological models of collective behavior, automatically discovers small groups of individuals traveling together in a low to medium density crowd (Figure 1). A pairwise distance that combines proximity and velocity cues is extended to form a robust distance between groups of people using a generalized, symmetric Hausdorff measure for inter-group closeness. Agglomeration of clusters is further constrained by an intra-group tightness measure inspired by sociological research into group behavior, enabling the number of groups in the scene to be determined automatically.

We validate our approach extensively on several video sequences taken in public pedestrian areas from elevated viewpoints typical of surveillance camera footage. Two indoor sequences are used to quantitatively compare results of our algorithm with consensus ground truth labeled by multiple human coders. We find that there is substantial statistical agreement between our algorithm's estimated groups and the human-perceived small group structure of the crowd. We also qualitatively evaluate our method on three outdoor sequences with different camera elevation angles, target sizes, and crowd densities, to demonstrate our method's tracking and group clustering capabilities across a range of conditions.

Although the experiments show that grouping based on trajectory distances is adequate to discover many small groups, two people walking side-by-side are more likely to form a group than two people who maintain a similar distance and speed but in a front-to-back configuration. We therefore hypothesize that the geometric layout of individuals will ultimately provide an important, and complementary, grouping cue. To that end, we also present a preliminary statistical analysis of the spatial configurations of small walking groups, using real-world pedestrian crowd footage with human annotations.

W. Ge is with the Computer Vision Lab, GE Global Research, Niskayuna, NY, 12309.

R.T. Collins is with the Department of Computer Science and Engineering, Penn State University, University Park, PA, 16802.

R. B. Ruback is with the Department of Sociology, Penn State University, University Park, PA, 16802.

This research was supported by the National Science Foundation's Human and Social Dynamics Program, Grant No. 0729363.

Analyzing the group structure of crowds has important practical applications. Current models of evacuation treat all people as separate agents making independent decisions. These "particle flow" models tend to underestimate the time it takes for people to leave an area, because groups of individuals who are together try to leave together, limiting the speed of the group to that of its slowest member. A small group behavior model also suggests new strategies for police intervention during public disturbances. Rather than seeing an irrational homogeneous crowd, police should be looking at small groups, only a few of which might merit coercion. Our work also demonstrates that computer vision is a viable methodology for supporting sociological analysis, enabling collection of empirical data on real crowds faster and more thoroughly than previously possible.

II. Background and Related Work

This section explains why the composition of a crowd is important for modeling social behavior and reviews related computer vision work on crowd scene analysis.

Collective Behavior and Small Groups. *Collective behavior* is the generic term for the often extraordinary and dramatic actions of groups and of individuals in groups [3]. Models of collective behavior tend to be bimodal. At one extreme are models that consider the entire crowd as one entity. Scholars have assumed that crowds transform individuals, so that the resulting collective begins to exhibit a homogeneous "group mind" that is highly emotional and irrational [3]. At the other extreme are models treating everyone as independent members acting to maximize their own utility. For example, crowd behavior has been simulated by considering people as particles making local decisions based on the principle of least effort [4].

As with most dichotomies, the truth is likely to lie in the middle. One hypothesis is that crowds are composed primarily of small groups, defined as a "collection of individuals who have relations to one another that make them interdependent to some significant degree" [5]. Despite being intuitively reasonable, there has been surprisingly little work to validate this hypothesis. Johnson [6] argues that most crowds consist of small groups rather than isolated individuals (see also [7]). An unpublished study by McPhail found that 89% of people attending an event came with at least one other person, 52% with at least 2 others, 32% with at least 3 others, and that 94% of those coming with someone left with the people they came with [8].

Our work in this paper analyzes sets of trajectories to discover small pedestrian groups in a crowd. Much has been written in the surveillance literature about detecting and tracking moving objects [9]–[11]. Here, we cover only recent work that focuses on analysis of crowd scenes and the identification of group behavior.

People Detection and Tracking. There have been several papers concerned with detecting a crowd and estimating its size. Often, the crowd is treated as a multiscale [12] or dynamic [13] texture, and extracted features are used to classify how many people are present [14]. Some approaches derive area-based count estimates by using prior calibration to relate the location and size of an image region to the number of people the region could contain [15], [16]. Similar approaches have been taken in [17] using holistic properties and in [18] using corners. Other research in vision addresses high-level crowd flow analysis in a statistical sense. This work includes identifying locations of roads/paths and learning patterns of normal scene activity from large datasets of individual trajectories [19], corner feature trajectories [20] or optical flow [21], [22]. Although these techniques are sufficient to generate predictive macro models of crowd motion, they do not address the problem of identifying and tracking groups of individuals. Indeed, measuring global crowd flow does not even require segmentation of the scene into individuals. Some recent efforts in multi-target tracking in crowded scenes have appeared in [23]–[25].

Behavior Analysis. Behavior recognition involving interpreting sequences of actions of one person or interactions of two or three are commonly built upon Hidden Markov Models [26] or Dynamic Bayes Networks [27]. These approaches are typically limited to a small, known number of individuals, due to the combinatorics involved in the coupled interpretation of multiple time series. There is recent evidence that more efficient recognition of group activities is possible by using a model of the group activity process to guide interpretation of the actions of individual members [28], [29].

More relevant to our work is recognition of collective behavior involving an arbitrary number of actors, such as identifying small groups of people shopping together [30], locating queues waiting at vending machines [31], analyzing social interaction in small group conversations [32], and recognizing crowd formation and dispersal behaviors through statistical clustering of pairwise relational predicates [33]. Only recently has collective locomotion behavior been studied. In [34], pedestrians with similar velocity are grouped together to aid motion prediction for tracking. This is a pragmatic definition of group, not a social one, since people who are far apart are clustered together when they have a common velocity. A model of social pedestrian groups based on measurement of each individual's personal space is explored in [35]. Cupillard et.al [36] develop a tracker to parse individual trajectories of people walking in a group. The tracker consumes motion detection results and links detected moving regions into paths (possible trajectories of individuals), following the principle of Reid's Multi Hypothesis Tracking (MHT) [37]. Special heuristics for pruning existing paths and creating new ones are developed based on considering how a path relates to a group structure. However, results are only shown on videos with several people (2–5) walking in a metro station, and it is not clear whether this method is scalable to larger crowds considering the larger number of MHT hypotheses. Lau et.al. [38] cluster 3D data points from a laser range finder into groups of human-sized blobs and adapt MHT to directly track moving, merging and splitting groups over time. An interesting aspect of their approach is the use

of anthropologist Edward Hall's theory of *proxemics* [39] to define groups based on ranges of personal and social interaction distances.

Group Behavior Models. Collective locomotion behavior is also studied in the traffic analysis and crowd simulation community. Models describing traffic flow can be characterized at the macroscopic or the microscopic level [40]. Because macroscopic studies focus more on the space allocation for pedestrians in a facility than on the direct interaction between pedestrians, they are not as suitable for predicting pedestrian groups as for evacuation planning. Microscopic models consider pedestrians as individual agents, with the collective crowd dynamics emerging from the interaction between agents. For example, in the *social forces* model [41], the behavior of an individual is subject to long-range forces caused by other pedestrians and environmental components such as obstacles and preferred areas. Similar approaches are used for multi-target tracking [42], [43] and abnormal behavior detection [44] in crowds. Another example is the Cellular Automaton (CA) model where individuals move according to a preference matrix that specifies the probabilities for a particular walking direction and speed [45]. Time and state are discretized in CA models, making them amenable to high-performance crowd simulation. *Floor field* models were introduced to substitute individual agents' intelligence with a floor field that is modified by the pedestrians and in turn modifies their preference matrices [46]. The advantage of using the floor field is that it can turn long-range interactions into local forces. In [21], floor fields are estimated from visual data and used to aid target tracking in dense crowds.

Relatively few models have been validated using real-world observations and even fewer explicitly consider group modeling. In [47], pedestrian behaviors follow a discrete choice model (DCM) that encodes a prior on walking speed and direction. Later, the DCM model was augmented by a leader-follower model and a collision avoidance model, both of which capture the interaction between pedestrians within a spatial range [48]. The leader-follower model (flocking) is one of the most frequently implemented models in crowd simulation [49], [50]. Other than that, small groups are often absent in crowd simulations [51] even though they are everywhere in real-world observations. We are only aware of the following works that look into more general group behavior models. In [52], crowd dynamics are modeled by a two-level hierarchical structure: a high-level group behavior model and a low-level action model. In [53], intra-group and inter-group influence matrices are used to specify interaction among group members and between groups respectively. The model is able to generate group structures ranging from linear line formations to clusters by varying the values of the influence matrices. However, no real-world data was used to validate the model. A recent study shows that social interactions among group members generate typical group structures that influence crowd dynamics: at low crowd density, group members tend to walk side-by-side, and this line formation is bent forward into a V-shape pattern as the density increases [54].

III. Detecting and Tracking Individuals

There is no shortage of explanations for crowd behavior, but there is a shortage of explanations supported by empirical sociological research [55]. The few empirical studies that have analyzed video data of people in public spaces (e.g. [56], [57]) have required hundreds of person-hours to hand code just minutes of film, greatly limiting the amount and type of video that can be quantitatively analyzed. The use of automated computer vision methods therefore could represent a substantial methodological improvement. However, generating a reliable set of trajectories for people in crowded public spaces is a non-trivial task due to frequent occlusions and the presence of nearby *confusers*. In this section we describe an approach for pedestrian detection and tracking that is capable of producing reasonable trajectories in crowded scenes containing closely spaced people. Clustering these trajectories to hypothesize small pedestrian groups is presented in Section IV.

We combine a pedestrian detector, a particle filter tracker, and a multi-object data association algorithm to extract long-term trajectories of people passing through the scene. The detector is run frequently (at least once per second), and therefore, in addition to any new individuals entering the scene, people already being tracked are detected multiple times. For each detection, a particle filter tracker is instantiated to track that person through the next few seconds of video, yielding a short-term trajectory, or *tracklet*. The goal at this stage is to generate a set of overlapping tracklets for each person. For example, if detection is run every 20 frames, and a particle filter tracks each detection through the next 80 frames, at any one moment in time roughly four temporally overlapping trajectory fragments will be measuring the location of any given person in any given frame. A second phase of trajectory-to-fragment data association is then run to link and merge these multiple fragments into single, longer trajectories. Below we describe our detection and tracking approaches in more detail.

A. Detection

We employ two different detection strategies. For videos captured from high elevation/wide angle views where people are small, we tackle pedestrian detection as a "covering" problem. Individual pedestrians are detected by using Reversible Jump Markov Chain Monte Carlo (RJMCMC) to find a set of overlapping rectangles that best explain or "cover" the foreground pixels in a binary segmentation generated by adaptive background subtraction. This method is similar to that of [58]–[60] and is capable of extracting overlapping individuals in crowds up to moderate density.

For higher resolution videos, pedestrian detection is performed in each frame using a combination of motion and contour (edge gradient) information, using a set of templates learned offline from training examples extracted from the same camera viewpoint. We use the HoG detector, implemented from the description in Dalal and Triggs [61]. Motivated by [62], we use motion information to determine regions that are more likely to contain moving pedestrians,

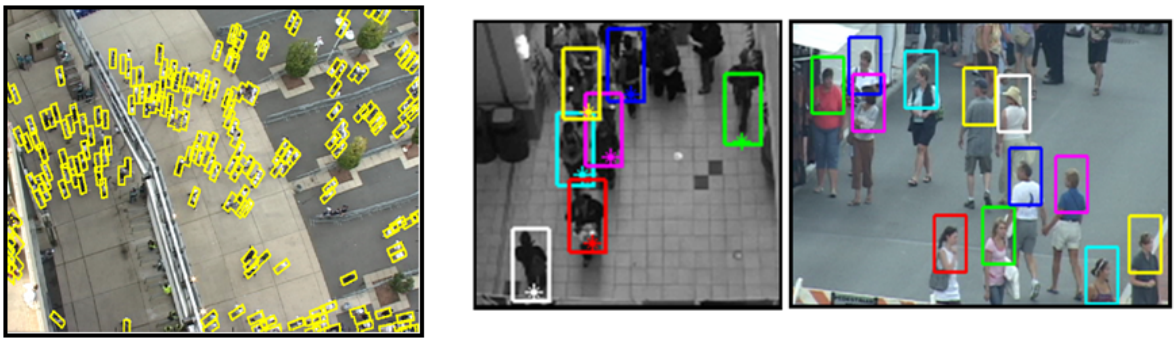

Fig. 2: Left: Sample detections in low resolution video by estimating a rectangular covering using RJMCMC. Right: Sample detections in higher-resolution video using an HoG detector for body (left) and head-and-shoulders (right).

in the form of a background subtraction mask. Background subtraction is helpful for suppressing image gradients in stationary regions of background clutter, to avoid finding false positives in those areas. Sample detection results from both methods are shown in Figure 2.

B. Tracking

For each detected pedestrian, a Sampling Importance Resampling (SIR) particle filter [63] is instantiated as a short-term tracker to track the detected person for the next few seconds of video. The state space is four-dimensional (x, y, u, v), where (x, y) is the hypothesized image location of the object centroid and (u, v) is the interframe velocity. We use constant velocity motion prediction with a Gaussian noise model. Roughly 50 particles are propagated for each target. The likelihood measure for determining particle weights for resampling can vary depending on resolution and quality of the video, e.g. normalized correlation of greyscale intensity templates, or Earth Mover's Distance (EMD) on marginal R,B,G color histograms. Since we reinitialize tracking frequently, the short-term tracker does not need to consider appearance model updates.

Sets of tracklets extracted in overlapping sliding windows of time are combined into longer trajectories by recursively merging each new set of tracklets into an evolving set of trajectories, one window at a time, in a single forward scan. Given a set of existing long-term pedestrian trajectories, and a new set of tracklets from the next sliding window, we match up trajectories to tracklets through a process of data association. Specifically, if there are N trajectories and M new tracklets, we form an NxM affinity table where each element contains a score rating the affinity of one trajectory with one tracklet that overlaps it in time. The affinity measure is a combination of geometric and appearance terms: a measure of "continuity" computed by the average distance between corresponding locations in the area of temporal overlap and appearance similarity of the targets. We also augment the affinity table with one row and column of "slack variables" to take into account that a new tracklet may not correspond to any existing trajectory (trajectory birth), or that a trajectory may not have been corroborated by any tracklet (which eventually leads to trajectory death).

To find the best assignment of trajectories to tracklets from the affinity table, we solve the corresponding Linear Assignment Problem (LAP) using the Hungarian algorithm [64]. Matched trajectory-tracklet pairs in the LAP solution are merged to extend the trajectory. Tracklets that have no matching trajectory are used to start new trajectories. Trajectories for which there is no matching tracklet have their "health" decremented. When a trajectory's health drops to zero, it is terminated. Trajectories that still exist at the end of this stage become the new trajectory set for another round of data association with tracklets in the next sliding window, and so on. The result of this forward scan procedure is the merging of multiple overlapping tracklets into a set of longer individual trajectories. An actual merge between two contiguous trajectories is a simple average of spatial locations. When two noncontiguous trajectories are merged, the locations in the gap between the two are computed by linear interpolation.

IV. Identifying Small Groups

In this section we present a clustering approach that hypothesizes small groups traveling together using the notion of group "entitativity" [65], defined in terms of criteria from Gestalt psychology: common fate (same or interrelated outcomes), similarity (in appearance or behaviors), proximity, and pregnance (patterning). Given a set of automatically extracted pedestrian trajectories, we identify potential groups within a sliding time window using hierarchical clustering based on robust measures computed from the noisy trajectories.

Our automatic grouping algorithm is inspired by McPhail and Wohlstein [57], who present the only objective measure we know of in the social science literature to determine which people are traveling together through the scene. In [57], group membership is determined via a cascaded set of three tests: 1) any two people who are within 7 feet of each other and not separated by another individual are considered to be contiguous, and pass on to the next test; 2) any two contiguous people whose speeds are the same to within .5 feet per second are judged to have the same speed, and pass on to the next test; and 3) any two contiguous people traveling at the same speed whose directions of motion are the same to within 3 degrees are judged to have the same

direction. A group-expand procedure is also defined to test whether a new individual should be added to an existing group. Note that in [57], these tests are applied by human observers who analyze frames of video offline.

A. Measurements

Consider the trajectory of a person in the scene as a set of tuples (s, v, t), where s is the position vector of the tracked person's centroid (projected into the ground plane using a homography estimated offline) and v is the velocity vector at frame t. Let Γ be the temporal overlap of the trajectories between person i and j within a temporal window T. We extend McPhail and Wohlstein's frame-based test to an aggregated pairwise distance measure between two trajectories over time:

$$w_{ij}^t = \alpha\mathcal{N}(\|s_i^t - s_j^t\|) + (1-\alpha)\mathcal{N}(\|v_i^t - v_j^t\|) \quad (1)$$

$$\delta_t(i,j) = \begin{cases} 1 & \|s_i^t - s_j^t\| < \tau_s \;\&\; \|v_i^t - v_j^t\| < \tau_v \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$\rho_{ij} = \sum_t \delta_t(i,j) \quad (3)$$

$$w_{ij} = \frac{\sum_t w_{ij}^t}{\rho_{ij}|\Gamma|} \qquad \text{for } i \neq j \text{ and } t \in \Gamma \quad (4)$$

where $\mathcal{N}(\cdot)$ is a min-max normalization operator applied independently for each pair of trajectories to linearly scale their velocity and distance differences into the range $[0, 1]$. We use a weight $\alpha = 0.7$ to combine spatial proximity and velocity cues into a pairwise distance w_{ij}^t computed at each time frame t. For each pair of tracked individuals, we compute the average pairwise distance w_{ij} over all the frames within Γ and scale by the number of times ρ_{ij} that the spatial distance and velocity difference between person i and j are below the thresholds τ_s and τ_v. This favors grouping people walking close to each other with similar velocities for a long period of time. The temporal consistency imposed by this aggregated measure helps overcome tracking errors to get stable groups over time.

Instead of considering the speed and direction differences separately as in [57], we compute the norm of the velocity difference vector because it is more robust against noise. Moreover, two people engaged in a conversation will have small speed if they are standing still, but can possibly have large random oscillations in orientation. The vector difference comparison is still stable in this case, and satisfies our expectation that people with coordinated behaviors are likely to be grouped together (Figure 6).

The pairwise distance metric is extended to measure the inter-group closeness between two groups of people by a generalized, symmetric Hausdorff distance. Hausdorff distance is a popular distance metric for two finite sets, and has been used for shape matching and trajectory analysis [2]. Here we use a modified version to measure the locomotion similarity between two sets of people. More formally, the symmetric Hausdorff distance between group

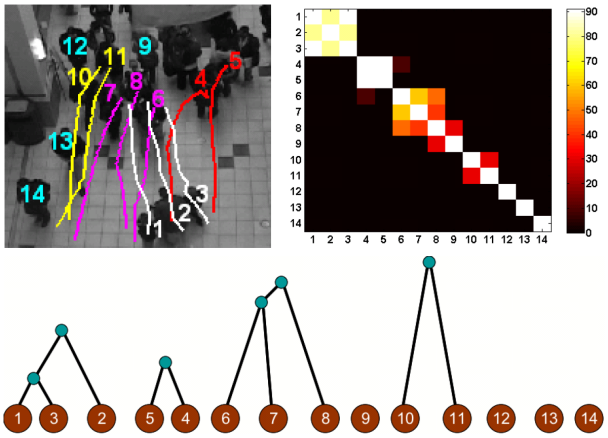

Fig. 3: Identifying small groups via agglomerative clustering. **Top** (left): Four groups (white, red, magenta, and yellow) are identified. (right): Pairwise counting value of ρ_{ij}. Brighter color indicates two individuals exhibit collective locomotion for a longer time; for example, node 9 can potentially be grouped with 8, but not 6 or 7. **Bottom**: Results of agglomerative hierarchical clustering.

A and B is $H(A,B) = \frac{h(A,B)+h(B,A)}{2}$, where

$$h(A,B) = \frac{\sum_{i=1}^{|A|}\sum_{j=1}^{\lceil |B|/2 \rceil} d_{il}}{|A| \times \lceil |B|/2 \rceil} \quad (5)$$

and d_{il} is the lth smallest distance amongst all the distances $w_{ij}, j \in B$, computed by Eqn.(4). The intuition behind this is that the directed distance from A to B is small when every member in A is close to at least half of the members in B, a rule also used in McPhail and Wohlstein's group-expand procedure.

B. Clustering

We identify groups based on a bottom-up hierarchical clustering approach that starts with individuals as separate clusters and gradually builds larger groups by merging two clusters with the strongest inter-group closeness (i.e., the smallest Hausdorff distance). Alternatively, one could take a top-down approach, starting with the entire crowd as a whole group and iteratively splitting into subgroups based on the same distance measure. We choose the bottom-up approach because it is more efficient in crowds composed of small groups.

Compared with other clustering methods (e.g. , K-means or spectral clustering), our approach does not require a predefined number of clusters. To automatically discover the number of groups, we construct a connectivity graph among people and measure the graph density as intra-group tightness. For any group of size $k \geq 1$, the vertices of the connectivity graph G_k correspond to the members in the group. There is an edge between vertex n_i and n_j iff person i and j are together for a sufficient amount of time, i.e. , $\rho_{ij} > \tau_t$ (Eqn. 3). We set $\tau_t = 10$ for all the experiments. The density of this graph helps us define intra-group tightness as follows. Let e_k be the total number of

edges in G_k and $\hat{e}_{k+1}$ be the minimal number of edges desired in G_{k+1} after including person p_i in G_k. Following the rule that a person i can be added to an existing group of size k iff she is connected with half of the existing group members [57], i.e. , the degree of $n_i \geq \lceil \frac{k}{2} \rceil$, we then have $\hat{e}_{k+1} = e_k + \lceil \frac{k}{2} \rceil$. By definition, $e_1 = \hat{e}_1 = 0$. For $k \geq 1$, given the basis condition that $\hat{e}_2 = 1$ and $\hat{e}_3 = 2$, we derive

$$\hat{e}_k = \begin{cases} (\frac{k}{2})^2 & \text{if k is even} \\ \frac{k-1}{2}(1 + \frac{k-1}{2}) & \text{if k is odd} \end{cases} \quad (6)$$

Two groups G_p and G_q then satisfy the intra-group tightness criterion if

$$e_{p+q} \geq \hat{e}_{p+q} + (e_p - \hat{e}_p + e_q - \hat{e}_q). \quad (7)$$

The terms in parentheses represent how many "extra" edges that group G_p and G_q can contribute to the intra-group tightness of the merged group G_{p+q}. A larger number of the terms in parentheses means that either or both subgroups are already tight groups with dense edges. In order for the merged group to remain a tight group, we require more edges in G_{p+q}, that is, a bigger number of e_{p+q}. Figure 3 illustrates how this tightness measure promotes the compactness of identified groups. Person 9 is excluded from the group $g = (6, 7, 8)$ because there is only one edge connecting 9 and 8, and including 9 in g does not satisfy the inequality specified in Eqn.(7). During each iteration of the merging process, we check the intra-group tightness of the next cluster to be merged. The clustering algorithm terminates when no clusters are qualified to be merged.

To summarize, within each temporal slice, starting from clusters with a single member, we gradually group people exhibiting collective locomotion by agglomerative hierarchical clustering. Each merging step is governed by both inter-group closeness, which is measured by a generalized, symmetric Hausdorff distance, and intra-group tightness, measured from the group connectivity graph. The latter provides a more principled way to determine when to stop clustering than manually setting a threshold.

V. Experimental Evaluation

We validate our proposed group detection method on a collection of videos of real-world pedestrian scenes with different environments (indoor and outdoor), viewpoints, pixels-on-target, and crowd sizes ranging from a few individuals to over 200. Sample video frames of each sequence are shown in Figure 4. Each video was recorded using a Sony DCR VX2000 digital video camcorder. After downloading the raw DV file from the tape, each video was converted to a sequence of PNG files using the open source program ffmpeg to produce deinterlaced 24-bit color images at a frame rate of 29.97 frames per second. We use full-body HoG detector on SU1, head-and-shoulders detector on ARTFEST , and RJMCMC rectangular covering detector on all the other sequences. For tracking, normalized correlation of greyscale intensity templates was used for SU1, which is a monochrome greyscale sequence; while color appearance likelihood (EMD on marginal R, B, G color histograms) was used in all other sequences. A summary of the group size statistics for sequences where ground truth was collected is listed in Table I.

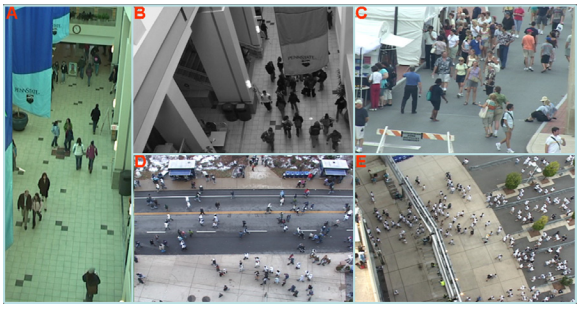

Fig. 4: Sample video frames of the test sequences used for the experiments in this section: SU2 (**A**), SU1 (**B**), ARTFEST (**C**), STADIUM1 (**D**), and STADIUM2 (**E**).

TABLE I: Percentage of groups of different sizes.

	1	2	3	4	5 or more
SU1 consensus	0.67	0.27	0.04	0.01	0.01
SU2 interview	0.64	0.32	0.02*	-	-
SU2 consensus	0.60	0.25	0.12	0.02	0.01

* This annotation only decomposes crowds into groups of size 1, 2, and 3 or more.

A. Data Collection and Annotation

Evaluation and comparison of work in this area is made difficult by lack of datasets with ground truth pedestrian groupings. We therefore have collected two datasets of pedestrians in a student union building from an elevated viewpoint and established "human consensus" ground truth by combining decisions made by multiple human coders.

The first experiment, SU1, was a pilot study performed on a four-minute video sequence. To obtain the ground truth, nine coders watched a version of the video with IDs overlaid on the 248 individuals passing through the scene. Coders were instructed to identify small groups by writing down the IDs of individuals in each group, and were told they could rewind and replay the video as often as needed. Groups determined by each coder were summarized into a numeric label for each pedestrian representing the size of the group they were traveling in (1 for single pedestrians, 2 for pairs, 3 for triplets, and so on). A *consensus ground truth* composite was computed by combining these numeric labels across all nine coders. The consensus label was defined as the modal response, the most common numeric label assigned by the coders. If there was more than one mode, the one with the smallest numeric label was assigned. Across coders, there was adequate, but not perfect agreement, which points out that there is some baseline ambiguity in deciding whether individuals form a group. For the 248 individuals in the video, all nine coders agreed about the coding of 161 individuals (65%), 6-8 coders agreed on the coding of an additional 57 individuals (23%), a bare majority of five coders agreed on the coding of 22 individuals (9%), and there was no consensus about 8 individuals (3%) (see Figure 5 Left).

Coders indicated that it was difficult to make judgments about groupings within the relatively narrow field of view

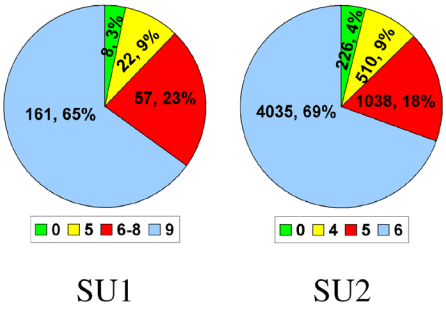

Fig. 5: Agreement rates among human coders for SU1 (left) and SU2 (right), reported in the form x, y% where x is the count of agreement cases and y is the percentage that count represents of the total cases.

of SU1. Based on their feedback, and to generate a longer and more challenging dataset, a second test sequence, SU2, was recorded. This sequence is one hour long, taken from a new viewpoint with a much larger range of depth, leading to more partial occlusion and a wider variation in image heights of people as they walk from near field to far field. The sequence also contains a wider range of crowd densities, from sparse (5 people per frame) to moderately dense (40 people per frame). Due to the length of the video, six coders were told to click on the heads of people in keyframes taken every 10 seconds, and to partition them into groups. Raw head clicks provided by all six coders were pooled to determine how many pedestrians were in each keyframe, and to assign each a unique ID and head location. As with SU1, group information provided by the coders was summarized by assigning each pedestrian a numeric label representing the size of group he or she was with. Of the 5908[1] pedestrians who were labeled in this way, all six coders agreed on the coding of 4035 of them (69%), five coders agreed on an additional 1038 (18%), a bare majority of four coders agreed on the coding of 510 individuals (9%) and there was no consensus about 226 people (4%), a similar rate of human coder agreement as in the shorter SU1 sequence, as visualized in Figure 5.

B. Quantitative Evaluation

SU1 sequence. We automatically detected and tracked pedestrians in the 4-minute SU1 sequence and applied hierarchical grouping to the generated trajectories to hypothesize small groups. Sample results are shown in Figure 6.

To quantitatively evaluate our grouping method, we first coded the consensus ground truth and computer-estimated group size for each pedestrian into one of two categories: alone or in a group. We achieved 89% agreement rate under this dichotomous coding scheme. Evaluating the results using a trichotomous coding scheme for each pedestrian (alone, in a group of two, or in a group of three or more), we achieved an 85% agreement rate. To test the statistical significance of the agreement between the computer estimates and the ground truth, we conducted Cohen's Kappa test on the trichotomous and dichotomous measures. In general, the κ score is defined as

$$\kappa = \frac{P_o - P_c}{1 - P_c}, \tag{8}$$

where P_o is the observed agreement among coders and P_c is the expected agreement if the coders had been making random decisions informed by the distribution of class labels, and thus agreeing purely by chance. Kappa scores range from -1 to 1, with the rate of agreement expected by chance yielding a score of 0. Similar to the Chi-squared test, Kappa measures agreement but also controls for the underlying base rates of the variables so that trivially predicting the group size that is dominant in the ground truth will not yield a good score. As was shown in Table I, the distribution of class labels for this application is not at all uniform, and therefore the conservative Kappa test is a more appropriate evaluation metric than agreement rate. Table II shows that there was substantial agreement ($\kappa > 0.6$) [66] between the consensus ground truth and the computer estimates.

TABLE II: Cohen's Kappa test on the indoor sequences.

	SU1		SU2	
	match rate	κ	match rate	κ
dichotomous	89%	.75	84%	.74
trichotomous	85%	.69	75%	.63

SU2 sequence. Similar experiments were conducted on the longer, more challenging SU2 sequence. We obtained similar though reduced match rates and kappa scores (Table II). Some sample detected small groups are shown in Figure 7. Besides the Cohen's Kappa test, we also computed the Adjusted Rand Index (ARI) [67], which is a standard statistical measure of the similarity in group membership between two set partitions, adjusted for chance in the same way that the Kappa test is. The ARI score is .65, which is again within the range of substantial agreement as measured on the Kappa scale. It shows that our method agrees well with ground truth not only on the size of groups, but also on the membership of the groups.

Further investigation on the SU2 sequence shows that the end-to-end performance of our approach degrades gradually as the crowd density increases, as shown in Figure 8. For a moderate crowd of 20 people per frame, our κ value is above .5, which still indicates reasonable agreement.

It is clear that grouping errors made by our algorithm are coupled with the underlying detection and tracking routines. Evaluation of our person detector alone shows a detection accuracy of 96% for detecting people in the ground truth keyframes with a false positive rate of 23%. Effects of tracking errors on grouping are harder to quantify. Our observation is that some tracking errors such as swapping identities between people traveling together do not affect

[1]Individuals appearing in multiple keyframes are labeled multiple times, therefore this number is larger than the number of unique individuals who passed through the scene.

Fig. 6: Sample group detection results in the SU1 sequence. Notice that the group marked with the rectangle has been consistently identified throughout a change of status from stationary to moving.

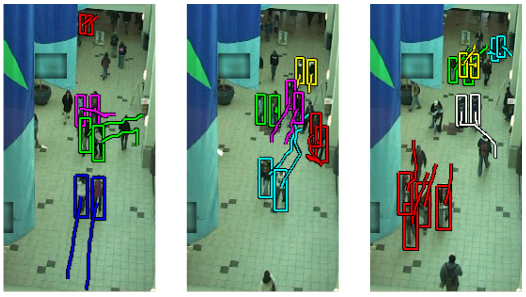

Fig. 7: Sample groups detected in the SU2 sequence.

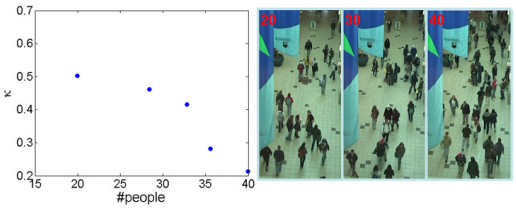

Fig. 8: Analysis of performance at higher crowd densities. **Left:** Dichotomous κ scores at various crowd densities. **Right:** Snapshots of crowds with 20, 30, and 40 people per frame (ground area is roughly 120 square meters).

the determination that they are a group, since their trajectories still overlap for a significant period of time. However, when the density of the crowd increases, the likelihood of a swapping error between people in different groups also increases, and these trajectory errors lead to errors in grouping.

To quantify how tracking errors are lowering the grouping performance, we further annotated thirty minutes of the SU2 sequence to obtain ground truth trajectories. The first half of the sequence (SU2-L) contains a relatively light crowd, while the second fifteen minutes of the sequence (SU2-H) contains a denser crowd. For each person in the sequence, hand labeling provided by human coders was processed to generate trajectory data points at every 10th frame. The statistical agreement between our computer-estimated groupings and human consensus ground truth improves (as expected) when using the ground truth trajectories, especially for the trichotomous scores, as shown in Table III.

TABLE III: Cohen's Kappa score of estimated pedestrian groupings on two portions of SU2, based on our method using hand-labeled trajectories, computer-estimated trajectories , and the method of Sugimara et.al. [68] using hand-labeled trajectories, from left to right in each column.

	dichotomous			trichotomous		
SU2-L	.88	.81	.62	.83	.66	.58
SU2-H	.75	.34	.33	.72	.27	.30

However, note that the improvement in performance is much higher for the higher density SU2-H sequence, which originally yields very low κ scores. We take this as an indication that our current tracker is not performing well during periods of high crowd density, and that improvement of this component will lead to a direct increase in better overall performance of the system.

C. Comparison with Corner Clustering

Theoretical Discussion. A number of works have considered the problem of counting pedestrians by clustering corner feature trajectories into groups that each represent a single individual [68]–[70]. These approaches can be viewed as potentially relevant to the problem here if one replaces the trajectories of corners with trajectories of people, and interprets the output clusters as representing small groups rather than individuals. Of these previous methods, Rabaud and Belongie [69] is the most closely related to our own. In that paper, hierarchical agglomerative clustering is also used to cluster trajectories into groups moving coherently through the image. Trajectories are first "conditioned" by spatial smoothing and extrapolation. A connectivity graph is formed with smoothed trajectories as nodes and edges linking pairs of trajectories that can be contained at all times within a box of expected object size, and that maintain a certain amount of rigidity in their pairwise distances. An initial clustering using RANSAC groups sets of four or more points that move together coherently with respect to an affine transformation. Finally, agglomerative clustering iteratively considers each closest pair of clusters in turn and merges them if all of their corner features are linked in the graph, stopping when all pairs of clusters have been considered.

In contrast, our small group detection is designed to be applied directly to noisy trajectories without first preconditioning them by smoothing. Rather than requiring full connectivity between all pairs of nodes when merging, our intra-group tightness criterion only requires each node of a cluster to be connected to roughly half of the nodes in the

other cluster, which is expected to be more robust when trajectories are noisy. Furthermore, rather than pruning an edge if the distance between two trajectories violates a threshold in *any* frame, our aggregated pairwise distance is based on "soft" measures that perform weighted averaging of spatial and velocity differences, normalized by the number of times both fall within a threshold. Therefore, distance thresholds can be violated in some frames without breaking the link between two trajectories. This not only yields better robustness to noisy trajectories, but in the present application allows us to find groups of people who came from different directions to meet up and then travel together, or who split up after a period of time to go their separate ways. The aggressive pruning scheme of [69] is fine for finding objects composed of many corner features, since even with a few dropped corners those objects will still be detected. In our application, however a typical small group is composed of only two or three people and the impact of false negatives is more damaging, motivating our softer approach to pruning.

Experimental Comparison. A second class of approach is represented by Brostow and Cipolla [70] and Sugimara et.al. [68]. These methods also form a connectivity graph, but do a one-shot pruning of edges based on a decision threshold, followed by finding groups as connected components in the remaining pruned graph. Here we describe Sugimara et.al. in more detail, which we have also implemented to perform a quantitative comparison. Edges in the initial connectivity graph are formed from a Delaunay triangulation of the x, y locations of people in the central frame of the sequence. Each edge between nodes p and q is assigned a weight representing a dissimilarity score, computed as the product of four numeric cues: 1) spatial proximity, 2) coherency of motion, 3) gait frequency, and 4) appearance consistency. Of these, only the first two make sense for our application (our nodes are trajectories of complete individuals, not corners on individuals) and are the same two features also used in Brostow and Cipolla. Spatial proximity is measured as the maximum distance between p and q in any frame, and motion coherency is the standard deviation of the distances between p and q over all frames in their period of temporal overlap. After computing the edge weights between trajectories in the initial connectivity graph, the graph is cut into connected subgraphs by simply pruning edges that have high dissimilarity scores, i.e. that have a weight above some threshold. As in [68], we set this threshold to half of the median value of all edge weights in the graph. After thresholding, each remaining subgraph is declared to be one group of individuals.

Results from Sugimara et.al. using hand-labeled trajectories for the SU2-L and SU2-H sequences are reported in Table III. There is a considerable drop in performance compared to our method using the same trajectories, particularly for the dense SU2-H sequence. The primary reason is that Sugimara's method produces overly-large groups from transitive chains of linked pairs that form a connected subgraph even though pairs near the beginning and end of the chain do not pass the test for being together in a group. The problem is that connected components of purely pairwise linkages cannot enforce higher-order consistency tests for group compactness, such as our intra-group tightness measure. A second drawback of all the corner trajectory clustering methods is an implicit assumption made about the statistics of intracluster vs intercluster edges in the connectivity graph. When there are 8-12 nodes (corners) forming each individual cluster, their $O(n^2)$ short pairwise edges represent a larger fraction of the total number of edges in a Delaunay triangulation [68] or pruned distance tree [70] than when there are only 2-3 nodes (people) per cluster. As such, many of the data-driven procedures in corner clustering approaches simply fail to work in our application. Even the affine motion check in [69] assumes more nodes per cluster than are typically present in a small pedestrian group.

D. Interview Study and Real-Time Observers

Although consensus judgments can be used as one estimate of whether individuals are walking together, the only way to know for sure whether individuals are walking alone or with others is to ask them. To collect this information, we had two research assistants briefly interview every fifth pedestrian while capturing the SU2 sequence. One research assistant, standing at the west end of the atrium, interviewed 78 pedestrians. The other research assistant, standing at the east end of the atrium, interviewed 68 pedestrians. Of the 181 individuals who were asked to be interviewed, 148 (81%) consented.

The interview consisted of two questions: (a) "Are you walking alone or with someone?" and (b) "If you are walking with someone, how many others are you walking with?" One assistant coded by hand the interview notes in the way described in Section V-A, using descriptive information in the notes to identify the respondents in the video keyframes. The 148 interviewed individuals yielded 419 total samples due to some individuals appearing in multiple key frames.

Furthermore, to determine how well human observers can spot pedestrian groups while watching a crowd in real-time, we had two researchers on site watch the pedestrians and write down the groups they observed. Observer A was positioned at ground level, and Observer B was viewing from an elevated location (the same location as the video camera). Written notes from these observers were also coded as samples in the manner described above.

Figure 9 shows a plot of pairwise statistical agreement tests (Cohen's κ scores) for the individual coders, real-time observers, and computer results, as measured with respect to both the interview ground truth and the human consensus ground truth. The raw κ values are also reported in Table IV for interview ground truth, and in Table V for human consensus ground truth. Since some sample sets such as the interview and real-time observer data are subsamples of the complete set of individuals, Cohen's Kappa agreement score between two sets is computed on the intersection of the samples measured in each set.

TABLE IV: Cohen's Kappa agreement scores based on interview ground truth for SU2.

	observer A	observer B	**computer**	coder 1	coder 2	coder 3	coder 4	coder 5	coder 6	consensus
dichotomous	0.22	0.56	**0.74**	0.86	0.85	0.75	0.82	0.85	0.82	0.85
trichotomous	0.29	0.54	**0.67**	0.81	0.79	0.63	0.82	0.79	0.83	0.83

TABLE V: Cohen's Kappa agreement scores based on human consensus ground truth for SU2.

	observer A	observer B	**computer**	coder 1	coder 2	coder 3	coder 4	coder 5	coder 6
dichotomous	0.5	0.7	**0.74**	0.92	0.93	0.85	0.88	0.92	0.92
trichotomous	0.53	0.71	**0.63**	0.91	0.93	0.83	0.88	0.91	0.92

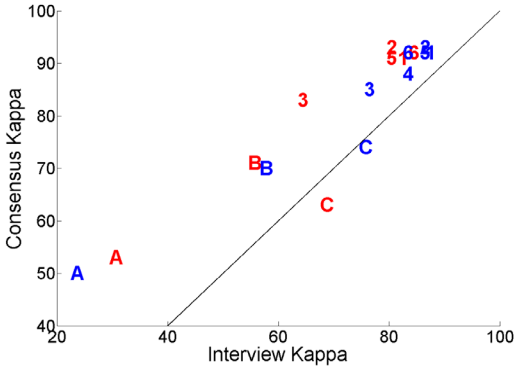

Fig. 9: Comparison of Kappa agreement values using interview and consensus ground truth. A and B are the two real-time observers. 1-6 are the six coders. C is the computer result. Dichotomous κ scores are shown in blue and trichotomous scores are in red. The solid line represents the curve where Consensus κ = Interview κ.

We see from Figure 9 that, among human decision makers, the ground-level real-time observer A performs the worst, real-time observer B with an elevated viewpoint does a better job, and off-line coders who can view and replay the video as long as they want perform the best. Computer performance is roughly on par with Observer B with respect to consensus ground truth, and exceeds Observer B with respect to interview ground truth. The computer results are also more comprehensive than the results from the real-time observers, in that 81% of the total pedestrian groups were accounted for by the computer, versus 68% for Observer B, and only 47% for Observer A. We therefore have an automated system with comparable accuracy to a real-time observer but with the advantage of being able to account for a greater percentage of individuals in the scene.

E. Qualitative Evaluation

In this section, we demonstrate our method qualitatively on three outdoor crowd sequences. The first two outdoor videos, STADIUM1 and STADIUM2, were captured during a sporting event. STADIUM1 is a five-minute clip taken of people walking on a closed street prior to the start of the game and STADIUM2 is a 30-minute clip taken of people leaving the stadium gate after the game. The camera was mounted on the stadium, thus the viewpoint is highly elevated and the image size of each person is relatively small. Figure 11(bottom) shows sample small groups found using our method.

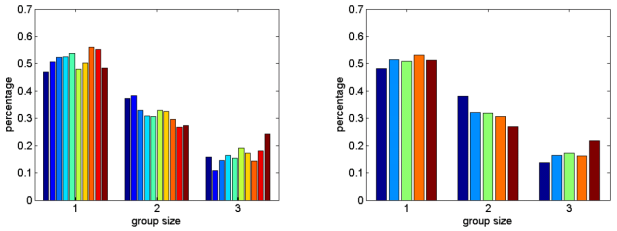

Fig. 10: Stability test on the STADIUM1 sequence. **Left**: the video was chopped into 10 segments, each of 1000 frames. **Right**: the video was chopped into 5 segments, each of 2000 frames. The estimated trichotomous coding of the small group structure of the crowd remains consistent across segments within each plot, and across both plots, suggesting that the grouping algorithm is consistent.

Ground truth evaluation of results from the STADIUM videos is difficult because human observers cannot make judgments easily with such large crowds. Instead, we evaluate the consistency of our grouping algorithm by chopping the video into segments, running the detection/tracking/grouping pipeline on each segment, and testing whether the estimated small group structure of the crowd remains stable over time. Two sets of experiments were conducted with different segment lengths. Figure 10 shows that the small group structure estimated by our algorithm remains consistent within each experiment and across experiments. However, this evaluation cannot rule out the possible presence of systematic bias in estimated group sizes.

The last outdoor video, the ARTFEST sequence, is a two-minute video captured at an outdoor art festival. The lower camera elevation angle, higher zoom, and "browsing" behavior of the crowd leads to frequent severe occlusion and more complicated trajectories. Figure 11(top) shows examples of detected small groups at different time frames where the crowd density varies and the trajectory pattern differs (e.g., strolling down the road vs pausing in front of a vendor).

The group parameters for all the test sequences are summarized in Table VI. For SU1 and SU2 with ground truth grouping information, the thresholds are decided by running a grid search to find values that maximize the kappa scores over a smaller training set. The parameters for other sequences without ground truth information were set empirically. The grouping results are not sensitive to small changes in threshold value.

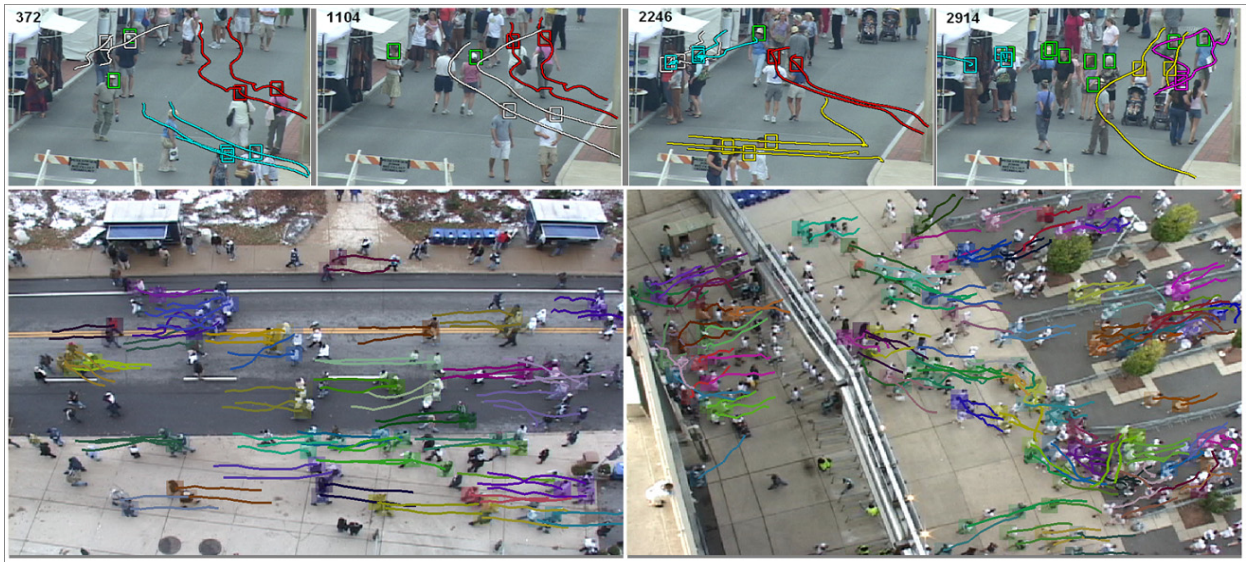

Fig. 11: Small group detections. **Top**: ARTFEST sequence. **Bottom**: STADIUM1 (left) and STADIUM2 (right) sequences. Trajectories of different groups are marked with different colors. The trajectories of people classified as traveling alone are omitted for clarity.

TABLE VI: Parameter Settings.

	τ_s	τ_v	τ_t
SU1	1.15 meters	0.01	10
SU2	1.06 meters	0.09	10
ARTFEST	115 pixels*	1.5	10
STADIUM1 and STADIUM2	30 pixels*	2	10

* Distance in ground plane is approximated by image distance when calibration information is not available.

VI. Group Configurations

In addition to identifying pedestrian groups, we are interested in understanding how groups move in crowds. The study of group behavior is not only of interest for sociologists but also of importance for realistic crowd simulation, evacuation planning, and vision tasks. For example, researchers have shown that pedestrian behavior models learned from video observations can be useful for tracking [21], [42], [43], [47] and activity recognition [44]. The prediction of pedestrian motion is usually determined by a repulsion force field, and it has been reported that a primary failure mode of this model is when groups of people walk together [42]. This finding suggests that a model component should be added to take group interactions into account. To build such a model, the first step is to study pedestrian walking patterns from real observations.

Discrete Choice Model. We have conducted a study on the general walking pattern of groups in normal crowds, i.e. not in evacuations or riots, using the SU2-L and SU2-H sequences, where human coders have annotated ground truth trajectories for individuals in the crowd. We also used the BIWI Walking Pedestrians dataset[2] containing two sequences (ETH and HOTEL) with similar ground truth annotations. From the ground truth trajectories and group information, we estimate individual behavior patterns using the discrete choice model (DCM) [47], where each individual moves according to a discretized set of choices based on his current speed and moving direction. There are three different radius zones, which correspond to deceleration, constant speed, and acceleration, defined as 0.5, 1.0, and 1.5 times the current speed, and eleven angular sectors that correspond to the discretization of the moving direction. Our analysis (not shown here) of DCM histograms aggregated over several groups over time indicates that, like individuals, members of small groups tend to maintain a constant walking speed and direction over short periods of time. However, as shown in Figure 12, individual DCM sequences alone are not adequate to model behavioral correlations between group members.

Statistical Shape Modeling. In order to study correlated patterns, we use a statistical shape analysis method [71] to analyze the spatial position of all group members jointly and estimate the typical group formations of walking pedestrians, which we refer to as *group configurations*. A group configuration S at a particular time consists of a point set of member locations, i.e. S is a $2g$-vector $[x_1, y_1, x_2, y_2, \ldots, x_g, y_g]^T$, where g is the group size. We first align each configuration with respect to its group center and moving direction so that each configuration S is centered at the origin and the group moving direction is aligned with the positive y-axis. After this alignment, we find the correspondence between the member points in different configurations by a data association procedure that minimizes the sum of squared distances between corresponding points in configurations from groups of the same size. Figure 13 shows the aligned configurations of groups of size three. We then stack all the aligned and matched

[2]http://www.vision.ee.ethz.ch/datasets/

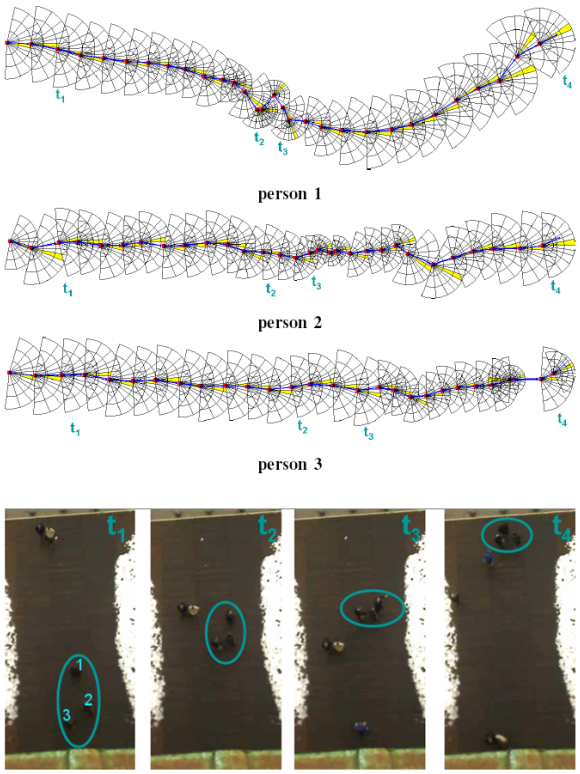

Fig. 12: DCM sequences from three people in a group. In each DCM, the current direction is colored in yellow and the next moving direction is colored in blue. The walking patterns exhibit strong correlations, e.g. person 1 stopped first to wait for the other two to catch up.

configurations column-wise into a $2g \times N$ sample matrix $\mathbf{S}$. Denote by $\hat{\mathbf{S}}$ the matrix $\mathbf{S}$ after centering (subtracting the mean $\bar{\mathbf{S}}$ from each column).

Principal component analysis is applied to the covariance matrix $H = \hat{\mathbf{S}}\hat{\mathbf{S}}'$ to study the joint variation in the group configuration samples. We model each configuration by

$$S \approx \bar{\mathbf{S}} + \mathbf{P}\mathbf{b}, \tag{9}$$

where $\bar{\mathbf{S}}$ is the mean configuration, $\mathbf{P}$ is the matrix of K dominant eigenvectors associated with the K largest eigenvalues of the covariance matrix, and $\mathbf{b}$ is a vector of K model parameters.

Results. In our experiment, the first four principal components account for most of the variability in the spatial configurations, explaining more than 99% of variations. In Figure 14 and in the accompanying video, we visualize these four common variations by varying $\mathbf{b}$. We name the four modes: deformation, stretching, rotation, and jittering.

For example, in Figure 14a, colored dots represent the mean shape and the blue lines indicate the variation along a particular principal component. We name it the deformation mode because the variation of the shape exhibits a deforming pattern: while the middle dot is moving down, the left and right dots are moving up, deforming from a straight line to a V-shape. Figure 14b shows the stretching mode. It is a pattern where the middle dot remains relatively still, whereas the left and right dots are moving in opposite

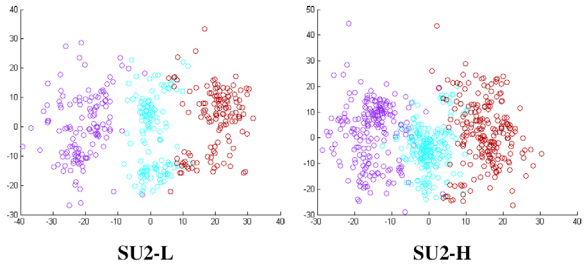

Fig. 13: The configurations of groups of size three are aligned with respect to their group centers and moving directions. The three members are plotted with three different colors after a data association procedure that matches points across different configurations. Edges indicating neighboring members of each group are omitted for clarity.

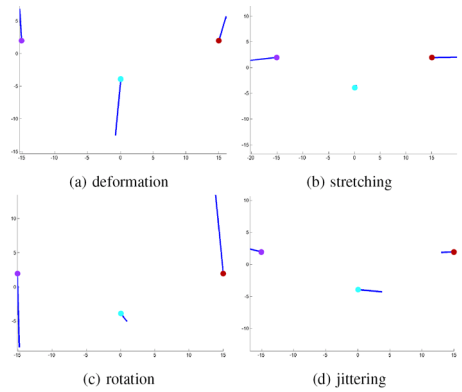

Fig. 14: The first four modes of variation on SU2, varying b from 0 to one std. See also the supplemental video.

directions, making the whole group configuration wider or narrower. Figure 14c shows the rotation mode, where the middle dot is the pivot point, and the left and right dots are rotating around the middle dot. Figure 14d shows the jittering mode, where the left and right dots are moving side-to-side in the same direction, opposite of the direction the middle dot is going.

Two of the four largest modes, *rotation* and *jittering*, may correspond to artifacts due to noise: jittering due to errors in the human-labeled trajectories and rotation due to noise introduced by the alignment process. The other two modes reveal interesting walking configurations. The *deformation* variation is the most interesting one, modeling the switch between different formations: V-shaped, line formation, and upside-down V-shaped. In [54], it is suggested that the V-shape structure is actively created and maintained by groups to facilitate social exchange. The upside-down V-shape is expected to be a more flexible structure than a line formation when a group of people are moving against an opposite traffic flow. Secondly, group members tend to maintain a distance from each other to keep the group formation, and they adjust their spread under various circumstances, e.g. to avoid environmental obstacles. The *stretching* mode defines this flexibility of group formations quantitatively.

These results motivate us to conduct further analysis in the future with a larger sample set of groups. For example, another possible explanation for the jittering mode is asymmetric offset of the center person with respect to the left and the right person. With more samples, we could conduct tests to distinguish this behavioral pattern from procedural artifacts.

VII. Conclusion

We have demonstrated that automated pedestrian detection and tracking can extract trajectories from video and that hierarchical clustering can detect small groups of people traveling together. To our knowledge, we are the first to show experimentally that results of agglomerative clustering are in substantial statistical agreement with subjective human perception of who is with whom in a crowd. As a field like computer vision matures, the importance of the research is measured in part by the influence it has on other fields. Our results demonstrate that automated tracking is capable of providing quantitative characterization of real crowds faster and with similar accuracy as human observation, providing a new methodology for the empirical study of social behavior. It is interesting to note that trajectory information alone is enough to yield substantial agreement with the perception of human coders who are able to address the grouping task by observing more subtle visual cues such as arm gestures and gaze direction.

Our future extensions include further investigation of small group configurations across different social events, which will be of interest to social studies of how pedestrians behave in different environments, and can be used for realistic crowd simulation. We also plan to explicitly incorporate the learned spatial group configurations to aid our tracking and grouping algorithms. For example, it would be interesting to apply the statistical model of spatial group configurations as a new social force feature for addressing the challenges of crowd tracking.

References

[1] A. Hoogs and A. G. A. Perera, "Video activity recognition in the real world," in *AAAI Conference on Artificial Intelligence*, Chicago, IL, 2008, pp. 1551–1554.

[2] X. Wang, K. Tieu, and E. Grimson, "Learning semantic scene models by trajectory analysis," in *European Conference on Computer Vision*, Graz, Austria, 2006, pp. 111–123.

[3] R. W. Brown, "Mass phenomena," in *Handbook of social psychology, Vol II*, G. Lindzey, Ed. Cambridge, MA: Addison Wesley, 1954, pp. 833–876.

[4] G. Still, "Crowd dynamics," 2000, ph.D. Thesis, University of Warwick.

[5] D. Cartwright and A. Zander, *Group dynamics: Research and theory (3rd. ed)*. New York: Harper, 1968.

[6] N. R. Johnson, "Panic at The Who Concert Stampede: An empirical assessment," *Social Problems*, vol. 34, pp. 362–373, 1987.

[7] A. Aveni, "The not-so-lonely crowd: Friendship groups in collective behavior," *Sociometry*, vol. 49, pp. 96–99, 1977.

[8] C. McPhail, "Withs across the life course of temporary sport gatherings," 2003, unpublished manuscript, University of Illinois.

[9] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans SMC-C*, vol. 34, no. 3, pp. 334–352, 8 2004.

[10] T. Moeslund, A. Hilton, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 103, no. 2-3, pp. 90–126, 11 2006.

[11] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L. Xu, "Crowd analysis: A survey," *Journal of Machine Vision and Applications*, vol. 19, no. 5-6, pp. 345–357, 10 2008.

[12] O. Arandjelovic, "Crowd detection from still images," in *British Machine Vision Conference*, Leeds, UK, 9 2008, pp. 1–8.

[13] A. B. Chan, Z. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, 2008, pp. 1–7.

[14] A. Marana, L. Costa, R. Lotufo, and S. Velastin, "On the efficacy of texture analysis for crowd monitoring," in *Proc. Computer Graphics, Image Processing and Vision*, Rio de Janeiro, Brazil, 1998, pp. 354–361.

[15] P. Kilamba, E. Ribnick, A. Joshi, O. Masoud, and N. Papanikolopoulos, "Estimating pedestrian counts in groups," *Computer Vision and Image Understanding*, vol. 110, no. 1, pp. 43–59, 4 2008.

[16] D. Kong, D. Gray, and H. Tao, "A viewpoint invariant approach for crowd counting," in *International Conference on Pattern Recognition*, Santa Cruz, CA, 2006, pp. 1187–1190.

[17] A. Chan, M. Morrow, and N. Vasconcelos, "Analysis of crowded scenes using holistic properties," in *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2009.

[18] A. Albiol, M. J. Silla, A. Albiol, and J. M. Mossi, "Video analysis using corner motion statistics," in *IEEE Workshop on Performance Evaluation of Tracking and Surveillance*, 2009, pp. 31–37.

[19] C. Stauffer and E. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions of Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[20] A. Cheriyadat and R. Radke, "Detecting dominant motions in dense crowds," *EEE Journal of Special Topics in Signal Processing*, vol. 2, no. 4, pp. 568–581, 8 2008.

[21] S. Ali and M. Shah, "Floor fields for tracking in high density crowd scenes," in *European Conference on Computer Vision*, Marseille, France, 10 2008, pp. 1–14.

[22] E. L. Andrade, S. Blunsden, and R. B. Fisher, "Modelling crowd scenes for event detection," in *International Conference on Pattern Recognition*, Hong Kong, 8 2006, pp. 175–178.

[23] M. Rodriguez, S. Ali, and T. Kanade, "Tracking in unstructured crowded scenes," in *IEEE International Conference on Computer Vision*, 2009.

[24] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybrid-boosted multi-target tracker for crowded scene," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[25] L. Kratz and K. Nishino, "Tracking with local spatio-temporal motion patterns in extremely crowded scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[26] N. Oliver, B. Rosario, and A. Pentland, "A bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 831–843, 8 2000.

[27] S. Gong and T. Xiang, "Recognition of group activities using a dynamic probabilistic network," in *IEEE International Conference on Computer Vision*, Nice, France, 10 2003, pp. 742–749.

[28] M. Ryoo and J. Aggarwal, "Recognition of high-level group activities based on activities of individual members," in *IEEE Workshop on Motion and Video Computing*, Breckenridge, CO, 1 2008, pp. 1–8.

[29] W. Zhang, F. Chen, , W. Xu, and Y. Du, "Hierarchical group process representation in multi-agent activity recognition," *Image Communication*, vol. 23, pp. 739–739, 1 2008.

[30] I. Haritaoglu and M. Flickner, "Detection and tracking of shopping groups in stores," in *IEEE Conference on Computer Vision and Pattern Recognition*, Kauai, HI, 12 2001, pp. 431–438.

[31] X. Naturel and J. Odobez, "Detecting queues at vending machines: A statistical layered approach," in *International Conference on Pattern Recognition*, Tampa, FL, 12 2008, pp. 1–4.

[32] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: a review," *Image and Vision Computing, Special Issue on Human Behavior*, vol. 27, pp. 1775–1787, 2009.

[33] A. Hoogs, S. Bush, G. Brooksby, A. Perera, M. Dausch, and N. Krahnstoever, "Detecting semantic group activities using relational clustering," in *IEEE Workshop on Motion and Video Computing*, Breckenridge, CO, 1 2008, pp. 1–8.

[34] A. French, A. Naeem, I. Dryden, and T. Pridmore, "Using social effects to guide tracking in complex scenes," in *IEEE Conf on Advanced Video and Signal Based Surveillance*, Hong Kong, 9 2007, pp. 212–217.

[35] J. J. Jr., A. Braun, J. Soldera, S. Musse, and C. Jung, "Understanding people motion in video sequences using voronoi diagrams," *Pattern Analysis and Applications*, vol. 10, pp. 321–332, 10 2007.

[36] F. Cupillard, F. Bremond, and M. Thonnat, "Tracking groups of people for video surveillance," in *European Workshop on Advanced Video-Based Surveillance System*, 2001.

[37] D. Reid, "An algorithm for tracking multiple targets," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 843 – 854, 1979.

[38] B. Lau, K. Arras, and W. Burgard, "Multi-model hypothesis group tracking and group size estimation," *International Journal of Social Robotics*, vol. 2, no. 1, pp. 19–30, 2010.

[39] E. T. Hall, "A system for the notation of proxemic behaviour," *American Anthropologist*, vol. 65, pp. 1003–1026.

[40] A. May, *Traffic flow fundamental*. New Jersey: Prentice Hall, 1990.

[41] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, pp. 4282–4286, 1995.

[42] P. Scovanner and M. Tappen, "Learning pedestrian dynamics from the real world," in *IEEE International Conference on Computer Vision*, 2009.

[43] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *IEEE International Conference on Computer Vision*, 2009.

[44] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[45] V. Blue and J. Adler, "Cellular automata microsimulation for modeling bi-directional pedestrian walkways," *Transportation research B*, vol. 35, no. 3, pp. 293 – 312, 2001.

[46] A. Schadschneider, "Cellular automaton approach to pedestrian dynamics - theory," *Pedestrian and Evacuation Dynamics*, pp. 75 – 86, 2002.

[47] G. Antonini, S. V. Martinez, M. Bierlaire, and J. P. Thiran, "Behavioral priors for detection and tracking of pedestrians in video sequences," *International Journal of Computer Vision*, vol. 69, no. 2, pp. 159 – 180, 2006.

[48] T. Robin, G. Antonini, M. Bierlaire, and J. Cruz, "Specification, estimation and validation of a pedestrian walking behavior model," *Transportation research part B*, vol. 43, pp. 36 – 56, 2009.

[49] S. R. Musse and D. Thalmann, "A model of human crowd behavior: Group inter-relationship and collision detection analysis," in *Workshop Computer Animation and Simulation of Eurographics*, 1997, pp. 39–52.

[50] M. Anderson, E. McDaniel, and S. Chenney, "Constrained animation of flocks," in *Eurographics/SIGGRAPH Symposium on Computer Animation*, 2003.

[51] B. Yersin, J. Maim, J. Pettre, and D. Thalmann, "Crowd patches: populating large-scale virtual environments for real-time applications," in *Symposium on Interactive 3D Graphics*, 2009.

[52] K. H. Lee, M. G. Choi, Q. Hong, and J. Lee, "Group behavior from video: A data-driven approach to crowd simulation," in *Symposium on Computer Animation*, 2007, pp. 109–118.

[53] F. Qiu and X. Hu, "Modeling group structure in pedestrian crowd simulation," *Simulation Modelling Practice and Theory*, vol. 18, no. 2, pp. 190–205, 2010.

[54] M. Moussaid, 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, 2010.

[55] C. McPhail, *The myth of the madding crowd*. New York: Aldine de Gruyter, 1991.

[56] W. White, *City: Rediscovering the center*. New York: Doubleday, 1998.

[57] C. McPhail and R. Wohlstein, "Using film to analyze pedestrian behavior," *Sociological Methods and Research*, vol. 10, pp. 347–375, 1982.

[58] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," in *IEEE Computer Vision and Pattern Recognition*, Madison, WI, 2003, pp. 459–466.

[59] G. M. Q. Yu and I. Cohen, "Multiple target tracking using spatio-temporal monte carlo markov chain data association," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[60] W. Ge and R. T. Collins, "Marked point processes for crowd counting," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[61] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Vision and Pattern Recognition*, San Diego, CA, 2005, pp. 886–893.

[62] P. Viola, M. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *International Conference on Computer Vision*, Nice, France, 2003, pp. 734–741.

[63] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: fifteen years later," in *In Handbook of Nonlinear Filtering*. University Press, 2009.

[64] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1955.

[65] D. Campbell, "Common fate, similarity, and other indices of the status of aggregates of persons as social entities," *Behavioral Science*, vol. 3, pp. 14–25, 1958.

[66] J. Landis and G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, 1977.

[67] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, pp. 846–850, 1971.

[68] D. Sugimura, K. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, "Using individuality to track individuals: Clustering individual trajectories in crowds using local appearance and frequency trait," in *International Conference on Computer Vision*, 2009, pp. 1467–1474.

[69] V. Rabaud and S. Belongie, "Counting crowded moving objects," in *IEEE Computer Vision and Pattern Recognition*, New York City, 2006, pp. 705–711.

[70] G. J. Brostow and R. Cipolla, "Unsupervised bayesian detection of independent motion in crowds," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 594–601.

[71] K. Mardia and I.L.Dryden, *Statistical Shape Analysis*. Chichester: Wiley, 1998.

Weina Ge is a computer scientist with the Computer Vision Group of GE Global Research Center. She received the Ph.D. degree in Computer Science and Engineering from The Pennsylvania State University in 2005. Her primary research interest is image and video understanding, especially tracking and behavioral analysis for pedestrian crowds.

Robert T. Collins received the Ph.D. degree in Computer Science from the University of Massachusetts at Amherst in 1993. He is an associate professor in the Computer Science and Engineering Department at The Pennsylvania State University. His research interests include video scene understanding, automated surveillance, human activity modeling, and real-time tracking. He is a senior member of the IEEE and a member of the IEEE Computer Society.

R. Barry Ruback is Professor of Crime, Law, and Justice and Sociology at Penn State University. He received a B.A. in history from Yale University, a J.D. from the University of Texas School of Law, and a Ph.D. in social psychology from the University of Pittsburgh. He is interested in collective behavior and is a consultant to the Pennsylvania Commission on Sentencing.