# Automatically Detecting the Small Group Structure of a Crowd

Weina Ge, Robert T. Collins, Barry Ruback
The Pennsylvania State University
University Park, PA 16802, USA
{ge, rcollins}@cse.psu.edu, bruback@psu.edu

## Abstract

*Recent work on computer vision analysis of crowds tends to focus on robustly tracking individuals through the crowd or on analyzing the overall pattern of flow. Our work seeks a deeper analysis of social behavior by identifying the small group structure of crowds, forming the basis for mid-level activity analysis at the granularity of human social groups. Building upon state-of-the-art algorithms for pedestrian detection and multi-object tracking, and inspired by social science models of human collective behavior, we automatically detect small groups of individuals who are traveling together. These groups are discovered using a bottom-up hierarchical clustering approach that compares sets of individuals based on 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.*

## 1. Introduction

There is an increasing interest in human behavior analysis from surveillance trajectory data, ranging from activity recognition based on the motion pattern of a single individual or the interactions among a few (e.g. [16]), to analysis of the flow of a large crowd, for example to discover pathways or monitor for abnormal events (e.g. [33]). 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 the small group structure of a crowd. Discovering small groups of people who are together provides a basis for further mid-level analysis of events involving social interactions of and between groups. It also has important practical applications in developing realistic crowd models/simulations for evacuation planning and real-time situation awareness during emergency response to public disturbances.

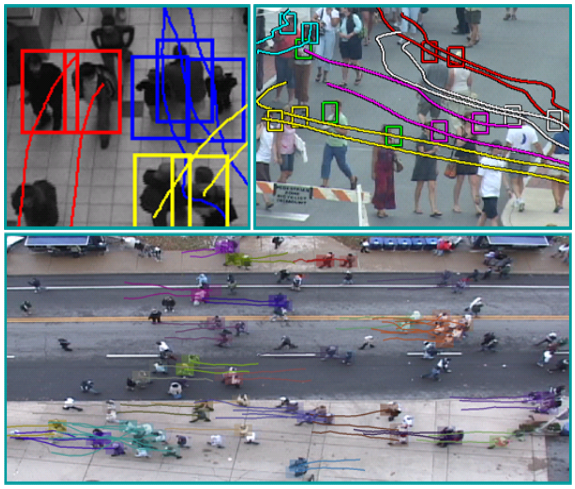

Figure 1. Small groups are prevalent in pedestrian scenes. Our algorithm detects groups of people traveling together via hierarchical clustering on trajectories automatically extracted from video of crowds under various conditions.

Our main contribution is development of a hierarchical clustering algorithm that, informed by sociological models of collective behavior, automatically discovers the small groups in a crowd (see Figure 1). A pairwise distance that combines proximity and velocity cues is extended to form a robust distance between groups (clusters) of people using a generalized, symmetric Hausdorff measure for inter-group closeness. Agglomeration of clusters is further constrained by intra-group tightness, a measure inspired by sociology research into group behavior, which enables us to automatically determine the number of groups in the scene.

We validate our approach on several video sequences taken in public pedestrian areas. 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 results and the human-perceived small group structure of the crowd. We further qualitatively evaluate our method on three outdoor sequences with differ-

ent camera elevation angles, resolution on target, and crowd densities, to demonstrate our method's tracking and group clustering capabilities across a range of conditions.

## 2. 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.

### 2.1. 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 [6]. 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 [6]. At the other extreme are models treating everyone as individuals. Under this view, a crowd is made up of 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 [32].

As with most dichotomies, the truth is likely to lie somewhere 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" ([8], p.46). Despite being intuitively reasonable, there has been surprisingly little work to validate this hypothesis. Johnson [18] argues that most crowds consist of small groups rather than isolated individuals.An unpublished study by McPhail found that 89% of people attending an event came with at least one other person.

From a "sociology meets computer vision" standpoint, Yan and Forsyth [35] apply vision techniques to automate analysis of the use of public spaces, in the spirit of sociologist William Whyte. Likewise, our work automates groundbreaking research by Clark McPhail, another pioneer in the use of video to analyze collective behavior.

### 2.2. Related Work on Vision-based Crowd Analysis

This paper is about discovery of small groups of individuals traveling together in a crowd from sets of trajectories. Much has been written in the surveillance literature about detecting and tracking moving objects to automatically extract trajectories, and we refer the reader to recent surveys in that area [17, 25, 37]. Here, we cover only relevant recent work that focuses on analysis of crowd scenes and the identification of group behavior.

There have been several papers concerned with detecting a crowd and estimating its size. Often, the crowd is treated as either a multiscale [3] or dynamic [9] texture, and extracted features are used to classify how many people are present [22]. 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 [20, 21]. 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 [31], corner feature trajectories [10, 5] or optical flow [1, 2]. 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.

Behavior recognition involving interpreting sequences of actions of one person or interactions of two or three are commonly built upon Hidden Markov Models [28] or Dynamic Bayes Networks [13]. 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 [30, 38].

More relevant to our work is recognition of collective behavior involving an arbitrary number of actors. Vision work addressing detection of collective behavior includes identifying small groups of people shopping together [14], locating queues waiting at vending machines [27], and recognizing crowd formation and dispersal behaviors through statistical clustering of pairwise relational predicates [15]. Only recently has collective locomotion behavior been studied. In [12], 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 well-motivated model of social pedestrian groups based on analytic measurement of each individual's personal space is explored in [19]. Social networks are discovered with the aid of face recognition in a Pan-Tilt-Zoom camera network in [36].

## 3. Identifying Small Groups

There is no shortage of explanations for crowd behavior, but there is a shortage of explanations supported by empirical research [23]. The few sociological studies that have analyzed video data of people in public spaces (*e.g.*, [34, 24]) 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 rep-

resent a substantial methodological improvement. Through user studies, we have observed that not only clicking the location of people is tedious, identifying who is traveling with whom is also difficult for human coders as the crowd density increases, simply because it is hard to keep track of each person in the scene.

In this paper we use vision-based detection, tracking and grouping algorithms to automatically identify small groups of pedestrians. Given a set of trajectories, our approach hypothesizes small groups traveling together using the notion of group "entitativity" [7], defined in terms of criteria from Gestalt psychology: common fate (same or interrelated outcomes), similarity (in appearance or behaviors), proximity, and pregnance (patterning). We chop the video into small temporal segments and identify all possible groups within a sliding time window by hierarchical clustering on robust measures computed from noisy trajectories. Since it is not always straightforward to observers whether an aggregation of individuals is a group as opposed to a mere collection of people [26], we design a clustering algorithm amenable to incorporating expert knowledge from sociologists.

Our automatic grouping algorithm is inspired by McPhail and Wohlstein [24], which is the only objective measure that we know of that has been put forth in the social science literature to determine which people are traveling together through the scene. In [24], two people are considered members of a group if they are within 7 feet of each other and not separated by another individual, have the same speed to within .5 feet per second, and are traveling in the same direction to within 3 degrees. A group-expand procedure is also defined to test whether a new individual should be added to an existing group.

### 3.1. Measurements

The trajectory of a person in the scene consists of a set of tuples $(s, v, t)$, where $s$ is the position vector of the tracked person's centroid and $v$ is the velocity vector at frame $t$. Let $\Gamma$ 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 people's trajectories over time:

$$w_{ij} = \frac{\sum_t w_{ij}^t}{\rho_{ij}|\Gamma|} \quad \text{for } i \neq j \text{ and } t \in \Gamma \tag{1}$$

$$w_{ij}^t = \alpha\mathcal{N}(\|s_i^t - s_j^t\|) + (1-\alpha)\mathcal{N}(\|v_i^t - v_j^t\|) \tag{2}$$

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

where $\mathcal{N}(\cdot)$ is a normalization operator that linearly scales data to $[0, 1]$, and $\delta_t(i, j)$ is set to 1 if $\|s_i^t - s_j^t\| < \tau_s \& \|v_i^t - v_j^t\| < \tau_v$ and 0 otherwise. We use a weighting parameter $\alpha$ to combine spatial proximity and velocity cues for the 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 time frames within $T$ and scale it with the number of times $\rho_{ij}$ that the spatial distance and velocity difference between person $i$ and $j$ are below the thresholds $\tau_s$ and $\tau_v$. This strategy favors grouping people walking close to each other with similar velocities for a long period of time. The aggregated measure yields robustness to tracking errors – although automatically extracted trajectories can be noisy, by considering temporal consistency we are still able to get stable groups over time.

Instead of considering the speed and direction differences separately, as in McPhail and Wohlstein, we compute the norm of the velocity difference vector because it is more robust against noise in the estimated trajectories. 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 4).

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 [33]. Here we use a modified version to measure the locomotion similarity between two sets of people. More formally, the symmetric Hausdorff distance between group $A$ and $B$ is $H(A, B) = \frac{h(A,B)+h(B,A)}{2}$, where

$$h(A,B) = \frac{\sum_{i=1}^{|A|}\sum_{j=1}^{\lceil |B|/2 \rceil} d_{ij}}{|A| \times \lceil |B|/2 \rceil} \tag{4}$$

and $d_{ij}$ is the $j$th smallest distance among all the distances between the person $i$ in $A$ and anyone in group $B$, computed by Eqn.(1) . 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 used in McPhail and Wohlstein's group-expand procedure.

### 3.2. Clustering

Mimicking the group-expand procedure of [24] where human coders iteratively check if an individual should be added to an existing group, we identify the groups based on a bottom-up hierarchical clustering approach that starts with individuals as separate clusters and gradually builds bigger 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. Consider the

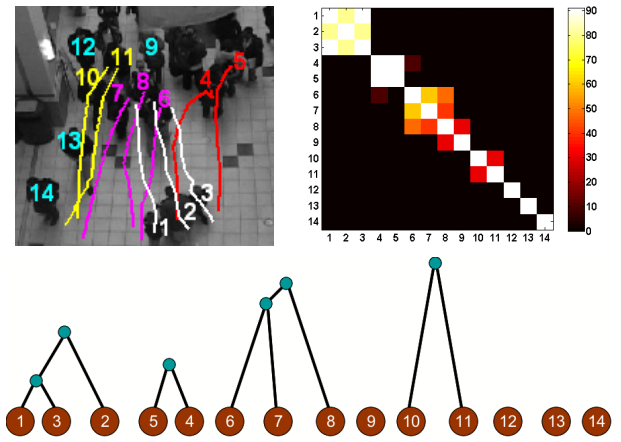


Figure 2. Identifying small groups via agglomerative hierarchical clustering. **Top** (left): Four groups (white, red, magenta, and yellow) were identified in a crowd. (right): Pairwise counting value of $\rho_{ij}$. Brighter color indicates two individuals exhibit collective locomotion for a longer time. **Bottom**: The result of agglomerative hierarchical clustering.

extreme case where all people in the scene are traveling alone – the bottom-up approach will terminate right away since there is no group to be built, whereas the top-down approach needs to work all the way down the hierarchy.

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 temporarily together for a sufficient amount of time, *i.e.*, $\rho_{ij} > \tau_t$ (Eqn.(3)). 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 [24], *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

$$\hat{e}_k = \begin{cases} (\frac{k}{2})^2 & \text{if k is even} \\ \frac{k-1}{2}(1 + \frac{k-1}{2}) & \text{if k is odd} \end{cases} \quad (5)$$

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

$$e_{p+q} \geq \hat{e}_{p+q} + (e_p - \hat{e}_p + e_q - \hat{e}_q). \quad (6)$$

Figure 2 illustrates how the 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.(6). 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. The latter provides a more principled way to determine when to stop clustering than manually setting a threshold.

## 4. Detecting and Tracking Individuals

The focus of this paper is our novel approach for clustering trajectories to hypothesize small groups of pedestrians using a social science model of human collective behavior. However, generating a reliable set of trajectories for people in crowded public spaces is itself a non-trivial task due to frequent occlusions and the presence of nearby *confusers*. Therefore, for completeness, we describe in this section our current approach for pedestrian detection and tracking, which is capable of producing reasonable trajectories in crowded scenes containing closely spaced people.

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, which is similar to that of [39], is capable of extracting overlapping individuals in crowds up to moderate density. Like all detectors, it produces both false postives and false negatives, and the subsequent tracking and grouping routines need to be robust to such errors.

Tracking individuals in the crowd is formulated as a multi-target tracking problem [4]. We use the Hungarian algorithm to perform multi-target data association between current trajectory hypotheses and detections in a new frame. The Hungarian algorithm finds an optimal bipartite matching between trajectories and detections given a table of pairwise affinities between likely match candidates based on proximity and similarity of appearance. As the set of current trajectories evolves, new ones are created from unmatched detections and old ones are removed if they have seen no supporting detections for a number of frames.

## 5. 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 densities ranging from a few individuals to over 100. Each video was recorded using a Sony DCR VX2000 digital video camcorder. After downloading the raw DV file from tape, each video was converted to a sequence of PNG files using the open source program ffmpeg to produce deinterlaced 24-bit color images of size 720x480 pixels at a frame rate of 29.97 frames per second.

### 5.1. Quantitative Evaluation

Evaluation and comparison of work in this area is made difficult by the lack of benchmark datasets with known ground-truth pedestrian groupings. We have collected two datasets of pedestrians in a student union building and established "human consensus" ground-truth by combining decisions made by multiple human coders. Recordings were taken from an elevated viewpoint to simulate typical surveillance video.

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 where numeric labels were overlaid on the 248 individuals who pass through the scene. Coders were instructed to identify small groups, and were told they could rewind and replay the video as often as needed. Group labels determined by each coder were translated into a numeric score for each pedestrian in the video (1 for single pedestrians, 2 for pairs, 3 for triplets, and so on). A "consensus ground-truth" composite score was computed by combining the labels from all nine coders. 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%).

Coders indicated that it was difficult to make judgments about groupings within the relatively narrow field of view. Based on their feedback, a second hour-long test sequence, SU2, was recorded from a new viewpoint with a much larger range of depth, which causes more partial occlusion and a wide variation in image heights of people as they walk from near field to far field. 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 (they were allowed to play the video forwards and backwards around each keyframe). Of the 5908 pedestrians who were labeled, 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 sequence.

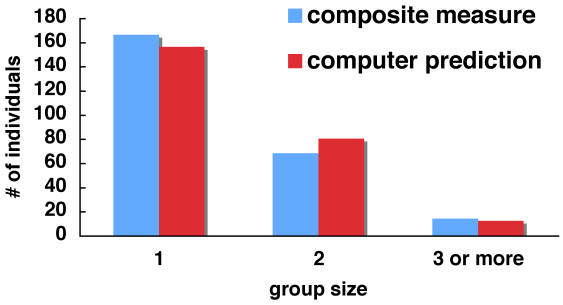


Figure 3. Trichotomous measure of small groups obtained from human coders compared with the computer predictions on the SU1 sequence.

**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 4. To quantitatively evaluate our grouping method, we first code ground-truth and computer predicted group size for each pedestrian into one of two categories: alone or in a group. We achieve 89% match rate under this dichotomous coding scheme. We also evaluated the results using a trichotomous coding scheme for each pedestrian: alone, in a group of two, or in a group of three or more (see Figure 3) with an 85% agreement rate. To test the statistical significance of the agreement between the computer estimates and the ground-truth, we conducted the Cohen's Kappa test on the trichotomous and dichotomous measures. Similar to the Chi-squared test, it 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. Table 1 shows that there was substantial agreement between the composite and the computer predictions, with Cohen's $\kappa > 0.6$.

| | SU1 | | SU2 | |
|---|---|---|---|---|
| | match rate | $\kappa$ | match rate | $\kappa$ |
| dichotomous | 89% | .75 | 88% | .74 |
| trichotomous | 85% | .69 | 77% | .62 |

Table 1. Statistical tests show substantial agreement ($\kappa > 0.6$) between human and computer group predictions for both the trichotomous and dichotomous cases on the indoor sequences.

**SU2 sequence.** Similar experiments were conducted on the first 15 minutes of the one hour SU2 sequence. Our results are again in good concordance with the human consensus ground-truth in the labeled keyframes (Table 1). Besides the Cohen's Kappa test, we also computed the Adjusted Rand Index (ARI) [29], which is a standard statistical measure of


Figure 4. Sample group detection results in the SU1 sequence. Links between people denote hypothesized small groups. Notice that the group marked with the rectangle has been consistently identified throughout a change of status from stationary to moving.

the similarity in group membership between two data clusterings/partitionings, adjusted for chance in the same way that 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 on the composition of the groups. Some sample detected small groups are shown in Figure 5.

Clearly the grouping error is coupled with the underlying detection and tracking routines. Evaluation of the our person detector alone shows an accuracy of $89\%$ for detecting people in the ground truth keyframes. Effects of tracking error on grouping are harder to quantify. Our observation is that some tracking errors such as swapping identities between people traveling together do not affect the determination that they are a group, since their trajectories still overlap for a significant period of time. We rely on an empirical investigation of the correlation between tracking and grouping as follows. For each set of trajectories of people in the frame, we set an upper bound on the maximal trajectory length by artificially shortening the long trajectories. Figure 6(left) shows that both the dichotomous and trichotomous match rates initially increase with the upper bound of the maximal trajectory length, then start to converge. For a range of trajectory length, the grouping performance stays stable.

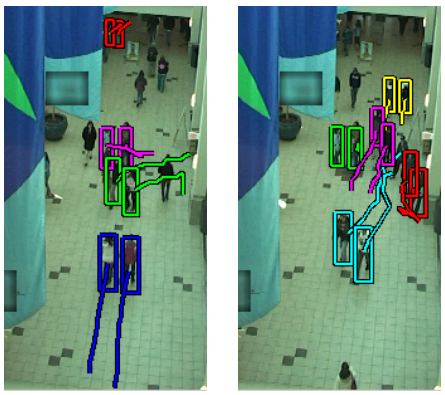

Figure 5. Sample small groups detected in the SU2 sequence.

### 5.2. Qualitative Evaluation

We have demonstrated that our end-to-end group detection results agree well with human consensus, which is a strong indication that either all pieces of our system are performing well, or else they are failing in ways that don't affect the final results. We also demonstrate our method 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 7(bottom) shows sample small groups found using our method. Evaluation in this case is even harder. Human observers cannot make judgements easily under such high crowd density. We design an empirical performance measure based on the observation that the small group structure of the crowd should stay stable over time. We evaluate the consistency of our grouping algorithm by chopping the video into segments and running the detection/tracking/grouping pipeline on each segment. Two sets of experiments were conducted with different segment lengths. Figure 6 shows that the small group structure estimated by our algorithm remains consistent within each experiment and across experiments. Notice the group structure is very different from the SU1 sequence, as shown in Figure 3, due to the social factor that people tend to go to sporting events in groups. Such difference in the composition of the crowd is worthy of further investigation for event recognition.

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. In this sequence, we performed frame-by-frame detection using a head-and-shoulders detector based on an SVM classifier trained on Histogram of Gradients (HoG) feature descriptors, inspired by the work of [11]. Figure 7(top) shows examples of

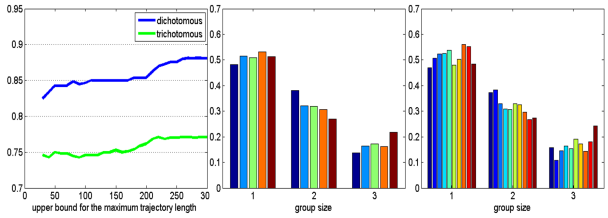


Figure 6. Performance analysis of the grouping algorithm. **Left**: The grouping performance is stable across a range of maximum trajectory lengths. **Middle**: stability test on the STADIUM1 sequence. The video was chopped into 10 segments, each of 1000 frames. **Right**: the video was chopped into 5 segments, each of 2000 frames. The estimated trichotomous coding of grouping results, indicating the small group structure of the crowd, remains consistent across segments within each plot and across both plots, indicating the grouping algorithm is consistent.

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).

## 6. 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.

**Acknowledgments.** This work was partially funded by the NSF under grant IIS-0729363.

## References

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

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

[3] O. Arandjelovic. Crowd detection from still images. In *British Machine Vision Conference*, pages 1–8, Leeds, UK, 9 2008.

[4] S. Blackman and R. Popoli. *Design and Analysis of Modern Tracking Systems*. Artech House Publishers, 1999.

[5] G. J. Brostow and R. Cipolla. Floor fields for tracking in high density crowd scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 594–601, New York, NY, 6 2006.

[6] R. W. Brown. Mass phenomena. In G. Lindzey, editor, *Handbook of social psychology, Vol II*, pages 833–876. Cambridge, MA: Addison Wesley, 1954.

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

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

[9] 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*, pages 1–7, Anchorage, AK, 2008.

[10] A. Cheriyadat and R. Radke. Detecting dominant motions in dense crowds. *EEE Journal of Special Topics in Signal Processing*, 2(4):568–581, 8 2008.

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

[12] 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*, pages 212–217, Hong Kong, 9 2007.

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

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

[15] 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*, pages 1–8, Breckenridge, CO, 1 2008.

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

[17] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Trans SMC-C*, 34(3):334–352, 8 2004.

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

[19] 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*, 10:321–332, 10 2007.

[20] P. Kilamba, E. Ribnick, A. Joshi, O. Masoud, and N. Papanikolopoulos. Estimating pedestrian counts in groups.


Figure 7. Small groups 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.

*Computer Vision and Image Understanding*, 110(1):43–59, 4 2008.

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

[22] 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*, pages 354–361, Rio de Janeiro, Brazil, 1998.

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

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

[25] T. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 103(2-3):90–126, 11 2006.

[26] R. Moreland and J. McMinn. Entitativity and social integration: Managing beliefs about the reality of groups. In V. Yzerbyt, C. M. Judd, and O. Corneille, editors, *The psychology of group perception: Perceived variability, entitativity, and essentialism*, pages 419–437. New York: Psychology Press, 2004.

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

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

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

[30] 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*, pages 1–8, Breckenridge, CO, 1 2008.

[31] C. Stauffer and E. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[32] G. Still. Crowd dynamics. Ph.D. Thesis, University of Warwick, 2000.

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

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

[35] W. Yan and D. Forsyth. Learning the behavior of users in a public space through video tracking. In *Workshop on Applications of Computer Vision*, pages I: 370–377, Breckenridge, CO, 2005.

[36] T. Yu, S. N. Lim, K. Patwardhan, and N. Krahnstoever. Monitoring, recognizing and discovering social networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[37] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L. Xu. Crowd analysis: A survey. *Journal of Machine Vision and Applications*, 19(5-6):345–357, 10 2008.

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

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