Social Group Discovery from Surveillance Videos: A Data-Driven Approach with Attention-Based Cues

Isarun Chamveha[1]
isarunc@iis.u-tokyo.ac.jp
Yusuke Sugano[1]
sugano@iis.u-tokyo.ac.jp
Yoichi Sato[1]
ysato@iis.u-tokyo.ac.jp
Akihiro Sugimoto[2]
sugimoto@nii.ac.jp

[1] The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan

[2] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

Abstract

This paper presents an approach to discover social groups in surveillance videos by incorporating attention-based cues to model group behaviors of pedestrians in videos. Group behaviors are modeled as a set of decision trees with the decisions being basic measurements based on position-based and attention-based cues. Rather than enforcing explicit models, we apply tree-based learning algorithms to implicitly construct the decision tree models. The experimental results demonstrate that incorporating attention-based cues significantly increased the estimation accuracy compared to the conventional approaches that used position-based cues alone.

1 Introduction

Vision-based sensing and understanding of human activities have been considered to be key techniques for security and marketing purposes. They include many research issues such as human detection, tracking, identification, path prediction, and action recognition, and these tasks are mutually related to one another. Knowing how pedestrians form social groups, *i.e.*, entities of two or more people with social interactions between them inside the same group, is of crucial importance among these issues to understand the scenarios in the video. For example, group information can be used to aid pedestrian tracking in low frame-rate videos [19, 24] and to analyze human behavior [6]. It also has the potential of being used for other tasks such as anomaly detection or path prediction.

For these reasons, techniques of social group discovery have recently attracted a great deal of interest, and several attempts have been made to discover social groups from videos. Ge *et al*. [14] proposed a method that aggregates pairwise spatial proximity and velocity cues and clusters them into groups based on the Hausdorff distance. Pellegrini *et al*.'s method [19] jointly estimated both pedestrian trajectories and their group relations by using third-order

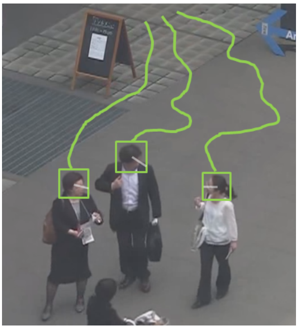


Figure 1: Example of ambiguous group relationship without head pose information. Squares and curved lines indicate tracked trajectories and lines in squares indicate their corresponding estimated head poses. Trajectories represented by same color indicate that they have been labeled as same group by annotator.

conditional random fields (CRFs) to model the relationships between them. Similarly, trajectories of individuals together with their groups are jointly estimated by applying decentralized particle filtering in an approach by Bazzani *et al.* [5]. Yamaguchi *et al.* [24] applied support vector machines (SVMs) with trajectory-based feature descriptors. Sochman and Hogg [21] proposed a method to infer social groups based on Social Force Model (SFM), which specifies several attractive and repulsive forces influencing each individual. A modified agglomerative clustering approach is then performed to infer pedestrian groups. Zanotto *et al.* [25] introduced an unsupervised approach based on an online inference process of Dirichlet Process Mixture Models.

These approaches demonstrate that position-based cues, such as the relative position and velocity of pedestrians, can be applied to solving the problem of social group discovery. However, attention-based cues, *i.e.*, how people pay attention to one another, have not yet been taken into account there. Attention-based cues have been utilized in several applications to analyze social interactions and these cues are proved to be effective in estimating human behaviors [1, 2, 13]. However, their applications to group discovery is not well studied. It is known that human attention is one of the most important cues for humans to distinguish social groups. For example, the attention of people in the same group tends to be focused on who is speaking during a conversation. Figure 1 is an example of a set of pedestrians in the same group during a conversation event. The relative distance between the rightmost person and the rest varies greatly over the trajectories in this example, this makes this case hard to be robustly estimated using position-based cues alone. However, the attention-based cues such as their eye gazes strongly suggest social groups, and can be used to help in this case.

In surveillance videos where high-level features such as eye positions cannot be obtained accurately due to low image quality, head poses can be a good approximation of human attention [3, 4, 18, 22]. Appearance-based head pose estimation from low-resolution images has been studied, and recent advances allows us to robustly infer the head poses of pedestrians from surveillance videos [17]. Indeed, recent works [8, 11, 12] report that head poses can be accurately estimated in real time even without using manually prepared training data.

We aim to combine attention-based cues and position-based cues to discover social groups in this work. This is the first work, to the best of our knowledge, to propose: 1)

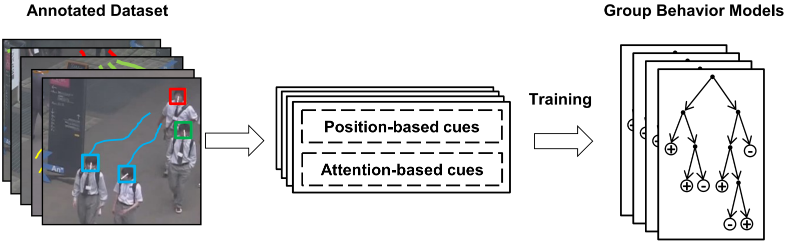


Figure 2: Given annotated dataset of pedestrian states, set of measurements based on attention-based and position-based cues is created for each pair of pedestrians. Estimator is then applied to construct set of decision trees representing group behavior models for social group discovery task.

a method that uses statistics of both attention-based and position-based cues over trajectories to discover social groups, and 2) a data-driven approach to find attentional behavior models for the group discovery task. Attention-based cues were explored in a recent approach by Bazzani *et al.* [6]. Their method imposes an explicit behavior model that pedestrians are likely to be in the same group if they are within other's view frustums and are standing sufficiently close to each other. However, especially when a history of cues over pedestrian trajectories is taken into account, there can be many other behavior models and finding an optimal model is not a trivial task. In contrast, we take a data-driven approach using both attention-based and position-based cues to building a classifier to detect social groups. We use a set of basic measurements obtained from such cues, and train a set of decision trees implicitly by using a supervised learning algorithm without enforcing explicit group behavior models.

2 Proposed Framework

Following Yamaguchi *et al.* [24], we define social group discovery as a pairwise problem to determine whether two people belong to the same group given information on their past states. Specifically, given a pair of past states $\{s_t^{(i)}\}$ and $\{s_t^{(j)}\}$ of pedestrians i and j, respectively, the goal is to assign binary label $y^{(i,j)} \in \{-1,+1\}$ that indicates whether they are in the same group $(+1)$ or not (-1).

The framework for our approach is outlined in Figure 2. We model the social group behaviors of pedestrians in a scene from two types of cues: *attention-based cues* and *position-based cues*. Attention-based cues are derived from observed human behaviors related to attentions, and position-based cues are derived from the observed trajectories of pedestrians.

Given a set of pair-wise pedestrian states $\{s_t^{(i)}\}$ and $\{s_t^{(j)}\}$, several measurements of both cues are calculated at each time step $t \in T^{(i,j)}$, where $T^{(i,j)}$ is a set of time at which both pedestrians i and j are observed and a collection of $|T^{(i,j)}|$ measurements are acquired for each measurement over the state pair. We aggregate the measurements into histograms to evaluate the frequencies of each behavior over the entire trajectory. Since not all histogram bins are informative for group discovery, we model social group behaviors as decision trees so that only informative histogram bins are used for decisions. Using an annotated dataset of pedestrian states, a tree-based learning algorithm is then applied to construct the decision

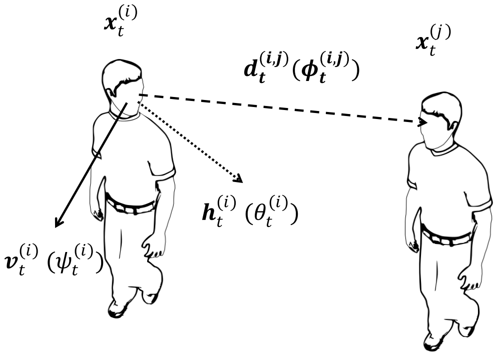

$\boldsymbol{x}_t^{(i)}$
$\boldsymbol{x}_t^{(j)}$
$\boldsymbol{d}_t^{(i,j)} (\boldsymbol{\phi}_t^{(i,j)})$
$\boldsymbol{h}_t^{(i)} (\theta_t^{(i)})$
$\boldsymbol{v}_t^{(i)} (\psi_t^{(i)})$

Figure 3: Relationships between pedestrian velocity $\boldsymbol{v}_t^{(i)}$, head pose $\boldsymbol{h}_t^{(i)}$, and displacement $\boldsymbol{d}_t^{(i,j)}$ and their corresponding image-plane angles $\theta_t^{(i)}$, $\psi_t^{(i)}$ and $\phi_t^{(i,j)}$, respectively. Brackets next to the vectors show their corresponding one dimensional angles.

trees representing group behavior models for the social group discovery task.

At each time step t, state variable $s_t^{(i)}$ representing pedestrian i is defined as $s_t^{(i)} = (\boldsymbol{x}_t^{(i)}, \boldsymbol{v}_t^{(i)}, \boldsymbol{h}_t^{(i)})$, where $\boldsymbol{x}_t^{(i)}$, $\boldsymbol{v}_t^{(i)}$ and $\boldsymbol{h}_t^{(i)}$ correspond to the position, the velocity, and the unit-length head direction of pedestrian i as illustrated in Figure 3. We denote image-plane angles of $\boldsymbol{h}_t^{(i)}$ and $\boldsymbol{v}_t^{(i)}$ as $\theta_t^{(i)}$ and $\psi_t^{(i)}$, respectively. $\phi_t^{(i,j)}$ is defined as an image-plane angle of the displacement vector $\boldsymbol{d}_t^{(i,j)} = \boldsymbol{x}_t^{(j)} - \boldsymbol{x}_t^{(i)}$. The angles are measured in radians.

2.1 Attention-Based Cues

Two types of attention-based cues are exploited in this work. The first cue is the *gaze exchange* between pedestrians. In order to perform group events, *e.g.*, conversation events, pedestrians in the same group often exchanged their gazes and fixed their attentions at one another. The second cue is the *mutual attention* of pedestrians in the same group. This is based on the observation that pedestrians often pay attention to the same object of interest. As was discussed earlier, we took an approach to learn the decision rules of these cues in a supervised manner using histograms of several measurements. This section introduces the details of the measurements, *i.e.*, the required building blocks to model these attention-based cues. For clarity, we omit the subscript t in what follows.

Difference in head pose and relative position. The first measurement is introduced to infer the *gaze exchange* cue. This measurement is defined as $a_1^{(i,j)} = |\theta^{(i)} - \phi^{(i,j)}|$, and calculates the degree to which pedestrian i directly looks at pedestrian j, which strongly indicates group events such as a group conversation or a group discussion.

Head pose difference. The second measurement, $a_2^{(i,j)} = |\theta^{(i)} - \theta^{(j)}|$, is intended to capture the *mutual attention* of pedestrians i and j. Since it is a difficult task to define objects of interest in every scene, we assumed that the objects of interest would be sufficiently distant from the pedestrians, *i.e.*, they shared mutual attention when the differences in their head poses were small.

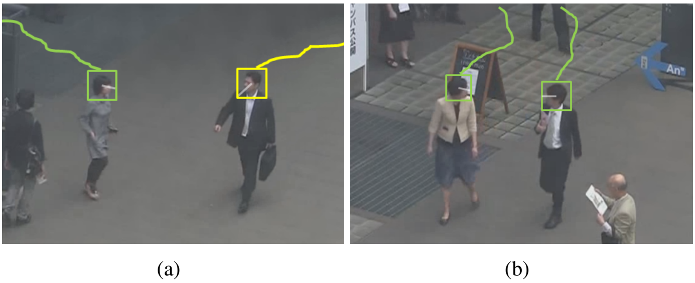


Figure 4: Examples of a_3 measurements. (a) An example when a_1 and a_3 measurements are low for both pedestrians. (b) An example when a_1 measurement is low and a_3 measurement is high for both pedestrians.

Difference in head pose and walking direction. While the previous two measurements are expected to capture attention-based cues, there are several different measures required for efficient decision models. For example, if pedestrians are walking, they naturally tend to look toward the direction they are walking. Therefore, looking toward the direction they are walking in does not suggest that pedestrians focus their attention on particular objects or other people in that direction, and the previous measurements are not necessarily informative. The third measurement is introduced to address such cases and is defined as $a_3^{(i)} = |\psi^{(i)} - \theta^{(i)}|$. This measures how steeply pedestrian i turns his head away from the direction he/she is walking in.

Figure 4(a) has an example when a_1 and a_3 are low for both pedestrians. Although both pedestrians are looking at each other, it is still ambiguous if they are in the same group. Figure 4(b) has an example when a_1 is low and a_3 is high for both pedestrians. The two pedestrians in this case are likely to be in the same group. With the walking focus measurements as decisions in the decision tree, our model can handle these two cases by taking into consideration a_1 and a_2 measurements only when a_3 measurements are sufficiently high.

Walking speed. The above assumption that pedestrians tend to look where they are walking does not hold for pedestrians walking slowly, *i.e.*, strolling or wandering around. The fourth measurement, $a_4^{(i)} = \|\boldsymbol{v}^{(i)}\|$, calculates the walking speed of each pedestrian, and is included to control the walking focus measurements.

2.2 Position-Based Cues

Measurements of position-based cues are derived from trajectories of two pedestrians based on the approach by Yamaguchi *et al.* [24] and are defined as follows.

- **Displacement:** $p_1^{(i,j)} = \|\boldsymbol{d}^{(i,j)}\|$ The distance between two pedestrians. Pedestrians in the same group tend to keep close to one another.

- **Difference in velocity:** $p_2^{(i,j)} = |\|\boldsymbol{v}^{(j)}\| - \|\boldsymbol{v}^{(i)}\||$ The difference in velocity between two pedestrians. Pedestrians in the same group tend to walk at the same speed.

- **Difference in direction:** $p_3^{(i,j)} = |\psi^{(i)} - \psi^{(j)}|$ The difference in direction between two pedestrians. Pedestrians in the same group tend to walk in the same direction.

- **Difference in direction and relative position:** $p_4^{(i,j)} = |\bar{\psi}^{(i,j)} - \phi^{(i,j)}|$ The angle between the average walking direction and the displacement vector between two pedestrians, where $\bar{\psi}^{(i,j)} = \frac{(\psi^{(i)}+\psi^{(j)})}{2}$ is the average walking direction of two pedestrians. Pedestrians in the same group tend to walk side-by-side, *i.e.*, in a direction perpendicular to their relative position.

- **Time overlap:** $p_5^{(i,j)} = \frac{|T^{(i)} \cap T^{(j)}|}{|T^{(i)} \cup T^{(j)}|}$. The length of overlapping time when pedestrians i and j appear on the scene up to time t, where $T^{(i)} = \{t' | t' \leq t, s_{t'}^{(i)} \neq \emptyset\}$ is a set of time steps where pedestrians i appear on the scene up to time t. Pedestrians in the same group tend to enter the scene at the same time.

2.3 Modeling of Social Behavior

To train decision trees, we calculate a set of measurements for every pair of pedestrians in the training set with the overlapping existent, *i.e.*, $\{(s_t^{(i)}, s_t^{(j)}) | t \in T_t^{(ij)}, T_t^{(ij)} \neq \emptyset\}$. As a_1, a_3 and a_4 measurements are measured for each pedestrian, two sets of the measurements are calculated. Therefore, at each time step, a total number of 7 measurements, $a_1^{(i,j)}$, $a_1^{(j,i)}$, $a_2^{(i,j)}$, $a_3^{(i)}$, $a_3^{(j)}$, $a_4^{(i)}$, and $a_4^{(j)}$ are collected for attention-based measurements. Each position-based measurement is calculated once and a total number of 5 measurements are collected at each time step.

We aggregate these measurements in a way that the decision rules of the model can be based on a single threshold, *e.g.*, how often people look at each other with less than τ_h degrees angles. Because comparing standard histogram bins imposes direct comparison of the measurement values, *e.g.*, how often people look at each other with the angles between τ_l and τ_h degrees, and is not always appropriate for aggregating attention-based measurements, we propose to use cumulative histograms. Each cumulative histogram is constructed with B_a equally-spaced bins. Histogram bins for a_1, a_2 and a_3 are placed between the range $[0, \pi]$. For a_4 measurement, we calculate the maximum speed v_{max} for pedestrians in the training set, and the histogram bins for a_4 are placed between the range $[0, v_{max}]$.

Position-based measurements are aggregated into standard histograms in the same manner as Yamaguchi *et al.* [24]. Each histogram is constructed with B_p equally-spaced bins. Histogram bins for p_1 are placed between the range $[0, d_{max}]$, where d_{max} is the diagonal length of the frame in the video. Histogram bins are placed between the range $[0, 2 \cdot v_{max}]$ for p_2 measurement, $[0, \pi]$ for p_3 and p_4 measurements, and $[0, 1]$ for p_5 measurement. Because training samples contain a different number of frames, both the standard and cumulative histograms are normalized so that total count in each histogram is summed to 1.

The decision trees could be implicitly constructed by using tree-based learning algorithms with the combination of histogram bins as decisions. A random trees classifier [10] was used in our approach to construct the trees. The histograms for each measurement were concatenated to represent the feature vector for that sample, and these vectors were used to train the random trees.

Table 1: Details of datasets used in our experiments. First three columns correspond to name, resolution, and duration of each sequence, respectively. Fourth column indicates number of trajectories annotated with group numbers. Fifth column indicates number of annotated groups for each sequence, and last column indicates average size of annotated groups in each sequence.

Sequence Name	**Resolution**	**Duration (minutes)**	**# of annotated trajectories**	**# of groups**	**Average group size**
UT-Surveillance	1920 × 1080	75	430	230	1.87
Town Centre	1920 × 1080	22	276	251	1.10

2.4 Pedestrian Tracking and Head Pose Estimates

We employ the head tracking method proposed by Benfold and Reid [7] to obtain walking trajectories of pedestrians in the video. Their method is based on a Kalman filter [15] with two types of measurements: the head locations given by a histogram of oriented gradient (HOG)-based head detector [20] and the velocity of head motion computed from multiple corner features [16, 23].

After pedestrian trajectories along with their corresponding head images are obtained, we apply the unsupervised approach proposed by Chamveha *et al.* [11] to obtain head poses. Their approach automatically aggregates labeled head images by inferring head direction labels from the walking direction. After outliers that were facing different directions had been rejected, their walking directions were used as ground truth labels of their head orientations. These ground truth labels were used to train the estimator for the task of estimating the head poses in our approach. Similar to [8, 11, 12], head directions in 2-D image plane are used as the approximation of actual head directions in our approach.

3 Experimental Results

We conducted experiments using two sequences: the UT-Surveillance sequence used by [11] and the Town Centre sequence used by [8]. The UT-Surveillance sequence contained pedestrians walking along a pathway, often in large groups. The Town Centre sequence contained pedestrians walking along a street. As the majority of pedestrians in this dataset walked individually, it contained many negative samples (pairs of pedestrians that did not belong to the same group).

Pedestrian trajectories and head poses were collected from the UT-surveillance dataset using the method mentioned in Section 2.4. For the Town Centre dataset, trajectories provided along with the dataset were used, and the head poses were obtained in the same way as the UT-Surveillance dataset. Tracked trajectories were manually annotated with social group IDs. Trajectories with erroneous or unstable results were ignored. The details of each dataset, the number of trajectories annotated with group numbers, the number of annotated groups, and the average size of groups are summarized in Table 1. Example frames in the sequences are shown in Figure 5.

We divided our annotated social groups into three disjoint sets and carried out three-fold cross-validation on the accuracy of estimates to evaluate the performance of our proposed method. Attention-based measurements were calculated and aggregated into cumulative his-

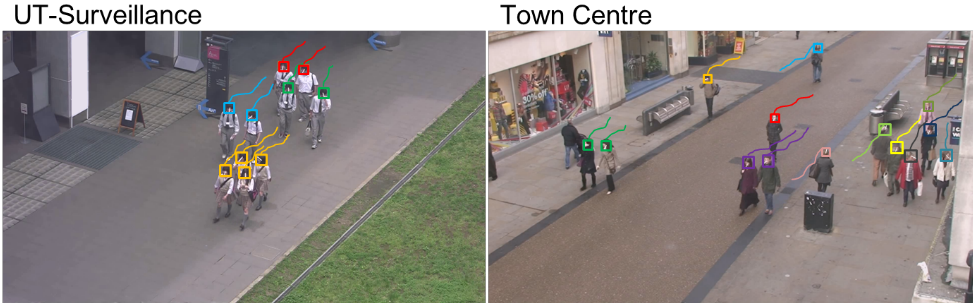


Figure 5: Sample frames from the sequences used in our work.

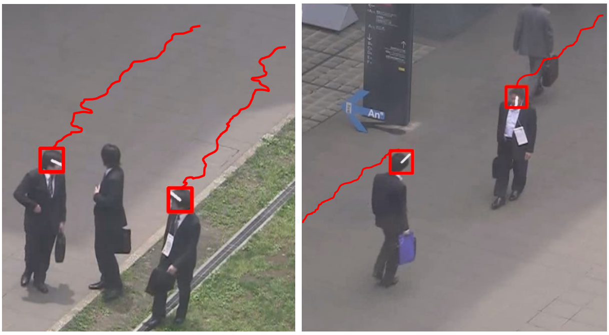
tograms with seven equally-spaced bins ($B_a = 7$), and position-based measurements were aggregated into histograms with seven equally-spaced bins ($B_p = 7$). These histograms were then concatenated as a 84-dimensional feature vector. The random trees is implemented using the OpenCV library [9] with the number of trees set to 400, the maximum depth of each tree set to 15, and the minimum samples in each leaf node set to 1% of total training samples.

We compared our method with the one by Yamaguchi *et al.* [24], who proposed to solve the same problem in the similar setting to our approach[1]. Comparisons were done by varying available past frames $N_{past} = 0$, 30, 60, 120, 240 and $N_{past} = \infty$. Measurements in these tests were calculated from at most N_{past} frames of each pair of pedestrians in the test set with overlapping time steps. We also conduct experiments on random trees trained with feature vectors obtained from position-based measurements alone to demonstrate the accuracy of random trees on position-based measurements.

Since the numbers of positive and negative samples were imbalanced in both data sets, balanced accuracy, *i.e.*, the average between the accuracy of each class, was used to evaluate the accuracy. The results are listed in Table 2. It can be seen that our approach improves the accuracy of social group discovery tasks in every case in both datasets. However, it can also be seen that with no past frame information ($N_{past} = 0$), our approach only slightly improves the accuracy, but with more available past frames, the improvements from Yamaguchi *et al.* [24] become more significant. These can be explained by the fact that attention-based cues are not always observed in every frame, *e.g.* pedestrians in the same group do not always talk to one another, and therefore the accuracy improvements were small in the cases with low N_{past}. However, attention-based cues can strongly suggest social group relationships even if such cues are rarely observed, *e.g.* the talking event is a strong indicator of social group, even if it occurs in a few frames. This makes accuracy improvements more significant with large N_{past}. In real application, however, long tracking trajectories are usually obtained from the tracker and the situation with low N_{past} are not typical. Therefore, high accuracy of the proposed approach can be expected in real situations. It can also be seen that the accuracy of the random trees classifiers trained with position-based cues is comparable to that by [24]. This shows that random trees is also an appropriate choice for position-based features.

We also measured the accuracy of our approach with the same settings as Yamaguchi *et al.* [24], who used low frame-rate videos. We tested our method with the datasets downsampled to 0.625 fps and the numbers of available past frames were $N_{past} = 0$, 1, 2, 4, 8

[1] We did not compare our method with [6] because [6] focuses on scenes where the individuals are stationary.



(a) Ground-truth: **+1**, Inferred: **+1** (b) Ground-truth: **+1**, Inferred: **-1**

Figure 6: (a) Example case where our method succeeded in inferring social group and (b) failed to infer social group. Same-group relationship between two pedestrians is correctly inferred in (a). Two pedestrians are inferred to be from different groups, while ground-truth stated otherwise in (b).

and $N_{past} = \infty$. The results are summarized in Table 3. Similar to the previous discussion, it is also the case in low resolution videos that our approach does not improve the accuracy on cases with small amount of available past frames N_{past}, but improves the accuracy with more available past frames. This shows that our approach is also applicable to low framerate videos, and can greatly improve the estimation accuracy given that some past frame information are available.

Figure 6(a) shows an example of a case where our approach correctly inferred that the two pedestrians were in the same social group. Even though they were walking at non-constant speed, the social groups were correctly inferred from attention-based cues. Our approach failed in inferring social groups, on the other hand, in cases where our assumptions about pedestrian behaviors do not hold. Figure 6(b) shows an example of such limitations. In this case, the two pedestrians in the same group are walking towards each other. It is our assumption that the *gaze exchange* cues are not informative when pedestrians turn their head in the direction they are walking in. Therefore, although the pedestrians are looking at each other, such information is disregarded and causes our approach to fail in this case. This suggests that more complex assumptions are needed to handle such case.

4 Conclusion

We proposed a data-driven method to discover social groups in surveillance videos by using attention-based and position-based cues. The introduction of attention-based cues allows complex relationships between pedestrians in the same group to be implicitly modeled as decision trees. The results from our experiments verified that our method improved the accuracy of social group discovery over an approach that used only position-based measurements. We believe that there are still other cues humans used to discover social groups, and investigating and discovering these cues will be important in future work.

Table 2: Accuracy of estimates with of our dataset. Accuracy is measured as average accuracy between two classes to avoid bias problems in the test samples. **Yamaguchi *et al.* [24]** presented results using the approach of [24]. **Trajectory + Random Trees** shows results using the features proposed by [24] and estimated using random trees estimator. **Proposed** indicates results with our proposed approach. Note that the frame rate in the datasets is 30 fps.

Dataset	Approach	N_{past}					
		0	30	60	120	240	∞
UT-Surveillance	Yamaguchi *et al.* [24]	74.9	75.8	76.5	75.8	76.0	76.4
	Trajectory + Random Trees	70.1	73.3	75.9	76.6	78.5	78.1
	Proposed	**76.4**	**76.6**	**77.9**	**77.9**	**80.3**	**81.2**
Town Centre	Yamaguchi *et al.* [24]	68.2	67.5	69.8	69.3	68.5	68.5
	Trajectory + Random Trees	67.3	67.3	68.8	70.1	70.1	70.3
	Proposed	**68.3**	**73.9**	**75.4**	**75.2**	**81.4**	**81.8**

Table 3: Accuracy of estimates using our dataset video downsampled to 0.625 fps.

Dataset	Approach	N_{past}					
		0	1	2	4	8	∞
UT-Surveillance (Downsampled)	Yamaguchi *et al.* [24]	75.1	75.3	75.2	75.3	75.8	75.8
	Proposed	**75.3**	**77.1**	**78.0**	**77.7**	**78.1**	**79.1**
Town Centre (Downsampled)	Yamaguchi *et al.* [24]	**67.0**	**67.6**	**69.1**	67.3	67.3	67.3
	Proposed	66.2	66.7	68.2	**72.7**	**71.2**	**72.7**

References

[1] Social interaction discovery by statistical analysis of f-formations.

[2] Sileye O. Ba and J.-M. Odobez. Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(1):101–116, 2011.

[3] S.O. Ba and J.-M. Odobez. Visual focus of attention estimation from head pose posterior probability distributions. In *Proc. 2008 IEEE International Conference on Multimedia and Expo*, pages 53 –56, 2008.

[4] S.O. Ba and J.-M. Odobez. Recognizing visual focus of attention from head pose in natural meetings. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(1):16 –33, 2009.

[5] L. Bazzani, M. Cristani, and V. Murino. Decentralized particle filter for joint individual-group tracking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1886–1893, 2012.

[6] L. Bazzani, M. Cristani, G. Paggetti, D. Tosato, G. Menegaz, and V. Murino. Analyzing groups: A social signaling perspective. In Caifeng Shan, Fatih Porikli, Tao Xiang, and Shaogang Gong, editors, *Video Analytics for Business Intelligence*, volume 409

of *Studies in Computational Intelligence*, pages 271–305. Springer Berlin Heidelberg, 2012.

[7] B. Benfold and I. Reid. Guiding visual surveillance by tracking human attention. In *Proc. 20th British Machine Vision Conference (BMVC)*, pages 14.1–14.11, 2009.

[8] B. Benfold and I. Reid. Unsupervised learning of a scene-specific coarse gaze estimator. In *Proc. the 13th IEEE International Conference on Computer Vision (ICCV)*, pages 2344 –2351, 2011.

[9] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[10] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[11] I. Chamveha, Y. Sugano, D. Sugimura, T. Siriteerakul, T. Okabe, Y. Sato, and A. Sugimoto. Appearance-based head pose estimation with scene-specific adaptation. In *Proc. 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1713–1720, 2011.

[12] C. Chen and J.-M. Odobez. We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video. In *Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1544–1551, 2012.

[13] Chih-Wei Chen, R.C. Ugarte, Chen Wu, and H. Aghajan. Discovering social interactions in real work environments. In *Proc. 2011 IEEE International Workshop on Social Behavior Analysis (SBA)*, pages 933 –938, 2011.

[14] W. Ge, R.T. Collins, and B. Ruback. Automatically detecting the small group structure of a crowd. In *Proc. 2009 Workshop on Applications of Computer Vision (WACV)*, pages 1–8, 2009.

[15] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[16] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. 7th International Joint Conference on Artificial intelligence*, volume 2, pages 674–679, 1981.

[17] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 31 (4):607 –626, 2009.

[18] K. Otsuka, J. Yamato, Y. Takemae, and H. Murase. Conversation scene analysis with dynamic bayesian network basedon visual head tracking. In *Proc. 2006 IEEE International Conference on Multimedia and Expo*, pages 949 –952, 2006.

[19] S. Pellegrini, A. Ess, and L. Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *Proc. 11th European Conference on Computer Vision (ECCV)*, pages 452–465, 2010.

[20] V. Prisacariu and I. Reid. fastHOG - a real-time GPU implementation of HOG. Technical Report 2310/09, Department of Engineering Science, Oxford University, 2009.

[21] J. Sochman and D.C. Hogg. Who knows who - inverting the social force model for finding groups. In *Proc. 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 830–837, 2011.

[22] R. Stiefelhagen. Tracking focus of attention in meetings. In *Proc. Fourth IEEE International Conference on Multimodal Interfaces.*, pages 273 – 280, 2002.

[23] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical report, CMU-CS-91-132, Carnegie Mellon University, 1991.

[24] K. Yamaguchi, A.C. Berg, L.E. Ortiz, and T.L. Berg. Who are you with and where are you going? In *Proc. 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1345–1352, 2011.

[25] M. Zanotto, L. Bazzani, M. Cristani, and V. Murino. Online bayesian non-parametrics for social group detection. In *Proc. 23th British Machine Vision Conference (BMVC)*, pages 1–12, 2012.