博士論文

Social Group Analysis from Surveillance Video using Attention-Based Cues

(人物の注意を手掛かりとしたサーベイランスカメラ映像における集団の解析)

東京大学大学院
情報理工学系研究科
電子情報学専攻

48-127409　ジャムウェイハー　イサラン

指導教員　佐藤 洋一　教授

Abstract

Understanding of social groups has recently attracted a lot of interest due to its application to various research fields. Social group cues have been used to improve the accuracy of pedestrian tracking in low frame rate videos, to analyze human behaviors, or to provide more realistic results for crowd simulation tasks.

It can be observed that members of social group tend to act in unison rather than individually, and activities performed in social groups also differ from those performed by individuals. Humans from the same social group tend to talk and pay attention to one another while ignoring those from different social groups, and social group members also tend to stay in close proximity and move in similar direction. These observations suggested that social group discovery can be performed with two types of visual cues, attention-based cues, such as how pedestrians pay attention to one another, and position-based cues, such as relative distances or movement direction between pedestrians.

This paper describes our approach to utilize these cues to discover social groups. Position-based cues are acquired through pedestrian tracking to collect spatial information of pedestrians in the video, while attention-based cues are calculated by analyzing various aspects of the attention of pedestrians in the video. The first part of this thesis describes an approach to estimate human attention. Using walking direction as a cue to infer head poses, our approach obtains human attention estimates without the need of

any prior training data. This approach also handles large variations that occur in head appearance within the same scene by segmenting a scene into multiple regions according to the similarity of head appearances.

A social group discovery approach is described in the second part of this thesis. We aim to improve social group behavior models by incorporating attention-based cues, which indicate social interactions such as conversation or discussion events of pedestrians in the video. Attention-based cues are modeled as a set of features based on human attention, and these features are used to learn a set of decision trees that represent the behaviors of social groups.

Finally, we propose an unsupervised approach to discover types of social groups by analyzing interactions among their members. A histogram of visual words is used to represent the characteristic of each social group. Social groups are then categorized by clustering these histograms. We also propose an approach to discover social groups that can change over time. Social groups are discovered separately in each frame with previous social group structures taken into account to help removing outliers.

This paper shows that social group information can be robustly acquired throughout these approaches. This would surely enrich current social group analysis processes and makes it possible for more detailed and real-time analyses, which would open up new possibilities for vast array of applications.

Acknowledgements

First of all, I would like to express my sincere gratitude to my professor Yoichi Sato for his kindness throughout my study throughout my Master and Doctoral study. Not only does he provide me with excellent advices and encouragements for my research, but he also help me with any problems I have while living in Japan. My life in Japan has been very smooth and enriched with wonderful experiences thanks to him.

I also would like to express my special thanks to professor Akihiro Sugimoto, Kris M. Kitani, Yusuke Sugano and Takahiro Okabe, who greatly helped me in my research and also for their advices. My research would not have been as successful without their guidances.

My thanks also go to my lab mates for their encouragements, advices and also their energetic personalities make me feel motivated to continue my research. I would like to express my appreciation towards my Thai friends in Japan, not only do they make me feel like home but also their helpful advices helped me get through a lot of problems in Japan.

I am also grateful to Sato laboratory's secretaries, Yoko Imagawa, Sakie Suzuki and Usui Chio who make my life here in Japan very smooth. Not only for processes related to my studies, but they also helped me with all of the processes I had while living in Japan. They always make complex processes seem very easy and simple to me.

My gratitude also goes toward the committees of my thesis pre-defense and defense sessions, professor Kiyoharu Aizawa, professor Shinichi Sato, professor Shunsuke Kamijo, professor Takeshi Oishi, and professor Toshihiko Yamasaki, for their invaluable advices toward the improvement of my thesis as well as their suggestions toward interesting future works.

I also expressed my gratitude towards Sato-Yo Scholarship. Not only the scholarship provides monetary aids and enables me to study here in Japan but also its numerous supports throughout my study here in Japan. The staffs extensively support my life in Japan and ensure that I live here without any problems. They always ask me whether I have any problems getting used to living in Japan or if I need any assistance. I also made a lot of friends from the scholarship, and I learned a lot about history and culture of various countries thanks to the activities provided by Sato-Yo scholarship.

Finally, my deepest thanks go to my parents for their love and their support throughout my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Vision-based sensing and understanding of human activities have been considered to be key components in techniques developed for many application domains such as security and marketing. Those components include human detection, tracking, identification, path prediction, and action recognition.

One aspect of human activity understanding is to comprehend how and what type of social groups human form. The definition of social groups has been studied since a long time ago. Some psychological studies define social groups in social cohesion context based on social interaction within groups [1, 2]. In this definition, members of social groups share some characteristics such as their interests, social backgrounds, or their common goal. Conversely, some works defined social group in cognitive context as “two or more individuals who share a common social identification of themselves” [3, 4]. These works perceive social groups as a unit where each member identify themselves as a member of the group. Social groups can also be categorized into primary groups which are formed by intimacy, such as the family, and secondary groups formed by the task-oriented needs, such as sports team and school students [5, 6].

Although it is difficult to directly measure psychological states of human mind, social group states of human can often be estimated through their social interactions, which is also a common aspect of social groups in the previous studies. It can be observed that members of social group tend to act in unison rather than individually, and activities performed in social groups also differ from those performed by individuals. Humans in the same social group move in similar direction and often engage themselves in conversation. Social interactions have been exploited in various computer vision literatures. In pedestrian tracking in low frame rate videos [7, 8], such information was used to improve ambiguous tracking results so that pedestrians in the same group are estimated to be near each other. Social group information was also used to aid human behavior analysis [9, 10] and even for tasks such as crowd simulation [11] to provide more realistic results.

In previous literatures, cues based on the spatial information such as position, walking direction and movement velocity of pedestrians in the video have been used for group discovery applications [8, 7, 12, 13, 14, 15]. Measurements of these position-based cues are derived based on the observation that pedestrians in the same social group tends to stay together and move in the same direction. However, human attention which is an important aspect in social interaction has been seldom taken into account in recent approaches.

Human attention is the process of concentrating on a discrete aspect of information while ignoring other perceivable information [16]. There are several types of attention such as tactile, auditory, and visual attention, and a lot of discussions have been made whether humans can have multiple attention at a time [17, 18]. While certain types of attention are extremely difficult to be estimated, visual focus of human is usually estimated as the attention of human in computer vision literatures and has served as a great cue to many applications. Recent advances have also shown that social behaviors such as group conversations or group discussions can be effectively estimated by using attention-based cues [19, 20] and suggests their great potential to solve the problem of social group discovery.

Attention-based cues are derived from human visual focus of attention, or where the person is looking at. Although visual focus of attention of human does not always coincide with their gaze such as when the person is turning his head during the attention shift [21], it is often accurate enough to estimate the attention by using their gazes in real scenario as is reported by several literatures [22, 23, 24]. However, stable gaze estimates are difficult to be obtained from surveillance videos where high-level features such as eye positions cannot be accurately obtained, and head pose estimates of humans are proved to be a sufficiently accurate estimation of human attention in such videos [25, 26, 27, 28], and as such is used in our work as an estimate of human attention for calculating the attention-based cues.

Social group type is also one of the interesting aspect of social groups which has a lot of potential applications. Supermarket owners might benefit from separating group of pedestrians standing near a shop stand and looking at the products from those who are passing without looking at it, in order to measure the attractiveness of the products. However, manual observation of such human behaviors in lengthy videos is time consuming and almost impossible to be conducted, and automatic detection and classification of such social groups are desired. Moreover, important social groups types are not always obvious. Pedestrian groups knowing each other slow down when they walk past, or those who are choosing what to buy in the convenience store will walk slowly near the stands they are interested in, and an unsupervised approach to discover these social group types is desired.

1.2 Related Works

We categorized the related works section into three parts. The first part describes approaches aimed at estimating human attention by means of head pose estimation, the second part describes state-of-the-art approaches on social group discovery, and the third part describes previous approaches related to social group types discovery.

1.2.1 Human attention estimation

Estimating human attention has always been an active topic in the field of computer vision studies. It is generally accepted that human attention coincides with their gaze, and several approaches have used human gazes to estimate their attention [22, 23, 24]. However, it is often difficult to do so in crowded scenes, and head poses are often regarded as acceptable estimates of human attention [25, 26, 27, 28]. Several approaches have proposed to estimate human attention in crowded scenes from head poses. Robertson *et al.* [29] used skin color as a descriptor and a binary tree algorithm to construct the head direction classifier. Body direction is also used to filter out poses that are not physically plausible. Benfold and Reid [30] proposed a descriptor that learns a model of skin color automatically and used randomized ferns for head direction estimation. Orozco *et al.* [31] proposed an image descriptor using similarity distance maps with class-mean appearance templates, and a multi-class support vector machine (SVM) for classifying head poses and estimating human attention. Benfold and Reid [32] acquired stable head tracking using a Kalman filter and estimated head poses using a randomized ferns classifier with the histogram of oriented gradients (HOG) features and color triplets comparisons (CTCs) as fern decisions to estimate human gazes and their attention. An approach by Schulz *et al.* [33] integrated pedestrian head localization and head pose estimation techniques to achieve highly accurate head pose estimates. Schulz and Stiefelhagen [34] trained eight head pose detectors to detect pedestrians' heads. The predictions were integrated over time to achieve improved robustness and efficiency.

These studies used head images with resolution as low as 20 × 20 pixels for training and testing the classifiers. While they are shown to work well for low resolution head images, they suffer from one important problem: a large number of training images with ground-truth labels, *i.e.*, correct head orientations, are required. Orozco *et al.* [31] used 800 manually cropped head images, 100 for each direction class from the i-LIDS [35] dataset. Gourier *et al.* [36] turned downsampled images from the Pointing'04 dataset into low resolution 23 × 30 dimension images. Robertson *et al.* [37, 29] used ground-truth samples produced by a human user drawing the line-of-sight of pedestrians in the images. Schulz *et al.* [33] used 7675 positive head pose samples and a set of negative non-head samples to construct the head pose classifier.

Unsupervised approaches have also been the focus of recent research on head pose estimation techniques to alleviate the cost of manual acquiring head pose training data. Benfold and Reid [38] constructed an unsupervised head pose estimator using a conditional random field model based on the same premise that people turn their heads toward where they are walking. An approach by Chen and Odobez [39] jointly estimated the body pose together with the head pose. This made it possible to filter out physically impossible head poses, which made their method more robust. Chamveha *et al.* [40] proposed an approach that automatically aggregated labeled head images by inferring head pose labels from the walking direction. Their approach also dealt with large variations that occurred in the appearance of heads within the same scene such as those caused by camera position and illumination differences.

1.2.2 Social group discovery

Social group discovery has recently attracted a great deal of interest, and several approaches have been proposed to discover social groups from surveillance videos. Position-based cues such as the relative position or walking direction of pedestrians have been used as the main measurements to discover social groups. Ge *et al.* [41] proposed a method that aggregated pairwise spatial proximity and velocity cues and clustered them into groups based on the Hausdorff distance. Trajectories of individuals together with their groups were jointly estimated by applying decentralized particle filtering in an approach by Bazzani *et al.* [13]. Sochman and Hogg [12] proposed a method of inferring social groups based on the social force model (SFM), which specifies several attractive and repulsive forces that influence each individual. A modified approach of agglomerative clustering was then applied to infer social groups. Zanotto *et al.* [14] introduced an unsupervised approach to discover social groups using pedestrians' positions, walking directions, and velocities with an online inference process of Dirichlet process mixture models.

Social group discovery approaches have also been used to aid pedestrian tracking in low frame-rate videos. Pellegrini *et al.*'s method [8] jointly estimated both pedestrians' trajectories and their group relations by using third-order conditional random fields (CRFs) based on human appearance and motion models. Yamaguchi *et al.* [7] proposed a behavioral model based on energy functions to predict the behaviors of pedestrians. Group information was modeled as one of the energy functions based on position-based cues and was calculated based on SVMs with trajectory-based feature descriptors. Group information served as an important cue to predicting the motion of pedestrians in a video. Qin and Shelton [15] proposed an approach to solving the problem with tracklet association. This approach was used to maximize the consistency of human appearances, human motions, and grouping cues and simultaneously solved the problems in both tracklet-tracklet association and tracklet-grouping assignment. Position-based cues were exploited in these approaches to discover social groups in videos. However, attention-based cues were not taken into account to discover social groups in any of their work.

1.2.3 Social group type discovery

Although there have been no research to directly estimate social group types, it can be considered a form of group activity recognition, which is one of the most active fields of research. Previous works on group activity recognition are categorized into several types. The first type considers group activities where each group member has its own role. Zhang *et al.* [42] used layered HMMs to recognize group activities in the meeting room. Dai *et al.* [43] proposed to recognize hierarchical structure of meeting using dynamic bayesian network (DBN). The second type of group activity recognition focuses on the motion of members in the group. Vaswani *et al.* [44] describes group activity as a shape which deforms over time, and abnormal activities are detected by comparing the shape polygons. Khan and Shah [45] fits a 3-D polygon to the group formation. The polygon is used to check whether the group is rigid or undergoing non-rigid deformation. Ryoo *et al.* [46] used description-based approach to recognize group. Their description can represent many group types due to the use of universal and existential quantifiers, specifying events that need to be performed by some or all of the members.

Unsupervised approach to group activity recognition has been done by Tang *et al.* [47]. Group activities in this work is defined as a group of people with similar activities in the video such as group of dancing people or people crossing the street. In this work, interest points extracted for each individual are used to learn a dictionary for each individual. Their dictionaries are then compared in order to find similar activities. Wang *et al.* [48] divided video into short clips and quantize local motion of each clip into visual words. The distribution of visual words are used to discover atomic activities, in which its distribution is used to discover interaction types in the video clip. These approaches, however, did not take into account interactions between human, which is the main components that constitute social groups.

1.3 Overview of this thesis

In this thesis, attention-based cues and their relations to social groups are thoroughly examined through series of experiments. We first described an approach to accurately estimate human attention in the video. Attention-based cues are then applied to improve the accuracy of social group discovery task, and we finally demonstrated that they can also help identifying social group types in an unsupervised manner.

First of all, Chapter 2 describes an unsupervised approach to head pose estimation. Human head pose is used as an approximation to their visual focus of attention, which is the basic component for attention-based cues. By exploiting the observation that human looks at where they are walking most of the time, we proposed a head pose estimation approach that do not require any prior train data and can also handle appearance differences within scene.

Chapter 3 describes our approach to group discovery. We aim to improve social group behavior models by incorporating attention-based cues, which indicate social interactions such as conversation or discussion events of pedestrians in the video. Attention-based cues are modeled as a set of features based on human attention, and these features are used to learn a set of decision trees that represent the behaviors of social groups. Thorough experiments are performed to demonstrate the effectiveness of the attention-based cues to the group discovery task.

In Chapter 4, an unsupervised approach to discover social group types is presented. Social group types are discovered by analyzing the interactions of pedestrians within the group. A set of visual words is constructed by clustering interaction vectors between pedestrians. A histogram of visual words is used to represent the characteristic of each social group. Group types are then identified by clustering these histograms. We also proposed an approach to discover social groups that can change over time. Social groups are discovered separately in each frame with previous social group structures taken into account to help removing outliers. Both quantitative and qualitative experiments were performed to evaluate the effectiveness of our approach.

Throughout the thesis, approaches to analyze human social groups are discussed as well as the effectiveness of attention-based cues used in the approach. Conclusions and possible directions for future works are given in Chapter 5.

Chapter 2

Unsupervised Head Pose Estimation with Scene Adaptation

In order to obtain visual focus of attention of human, one of the greatest cues is their head pose. Human head pose conveys a lot of information. Human looking at the pictures in the gallery often turn their head to the picture, while those having conversations would turn their head to the speaker. Therefore, in modern literatures, head pose of human is used as an approximation of their attention, and various researches reported to obtain highly accurate results [25, 26, 27, 28].

2.1 Introduction

Estimating the human visual focus of attention has recently become a popular research trend, as such research has numerous applications in our daily life. For example, it can be used to estimate the attention of people walking along the street [32, 49]. Having attention information enables us to easily infer interaction between people and to consequently analyze human interaction or identify on-going activities without requiring any human assistance [20, 19]. Head pose is known to be an important factor in inferring the focus of attention of humans. Therefore, techniques for estimating head pose are considered important and have attracted great interest recently.

Various image-based approaches have been proposed for estimating eye gaze. However, most of them[1] require high resolution images [51, 52] or special equipment such as depth cameras [53, 54] or actively controlled pan-tilt-zoom cameras [55]. However, head regions in in visual surveillance images are often quite small and therefore contain limited information. Accurately estimating head pose in such cases remains a challenging task.

The use of appearance-based approaches is thought to be promising for estimating head pose from low resolution images. Compared with model-based methods such as active appearance models [56, 57], which rely on geometric facial models and require localization of facial elements, appearance-based methods directly use pixel values of an image as an input to extract image features and are known to be effective even with low resolution images.

Appearance-based head pose estimation approaches rely heavily on a dataset used for training estimators. This is due to the fact that head appearances can change significantly from scene to scene. Even in the same scene, there could be substantial differences in head appearance due to extreme differences in illumination or camera viewing angle. Therefore, a training dataset is best taken from the same location as the target data. However, collecting ground-truth training samples is a labor-intensive and time-consuming task, and it is prohibitively expensive to collect ground-truth data manually every time a head pose estimation method is applied to different scenes.

We propose an appearance-based head pose estimation method in order

[1] We refer readers to [50] for a recent survey on approaches to head pose estimation.

to overcome these problems. Our method is based on two key ideas: automatically acquiring a training image dataset with ground-truth head pose and segmenting a scene into multiple regions with similar head appearances. We first construct a training image dataset by tracking pedestrians in a scene of interest to capture head images where their pose is regarded as a ground truth head orientation. To address the problem of appearance differences within a scene, the scene is segmented into multiple regions based on the similarity of head appearances, and a head pose estimator is then trained for each region. This approach enables us to test each head image with the estimator trained with data taken from the same region. Higher accuracy can thus be expected because the data used to train the estimator have a similar appearance to the test data.

2.2 Proposed Framework

Appearance-based head pose estimation involves determining a head pose p from a feature vector $\boldsymbol{h}$ of an input head image. We define p as the head pose in an image plane. In our work, head pose estimation is defined as a regression task, where head pose is defined as continuous angles as illustrated in Figure 2.1.

With a set of training samples $D = \{(\boldsymbol{h}_k, p_k)\}$, the mapping $p = f(\boldsymbol{h})$ between the head pose and the feature vector can be learned through various regression algorithms. The mapping function then can be used to estimate a head pose p^* from a new feature vector $\boldsymbol{h}^*$ in test scenes.

As discussed above, an important problem yet largely ignored in previous studies is how to obtain appropriate training samples D. Since we implicitly assume the mapping function $f(\boldsymbol{h})$ is identical in both training and test datasets, estimation accuracy highly depends on how similar the head images are in both datasets. Due to various factors such as lighting conditions or

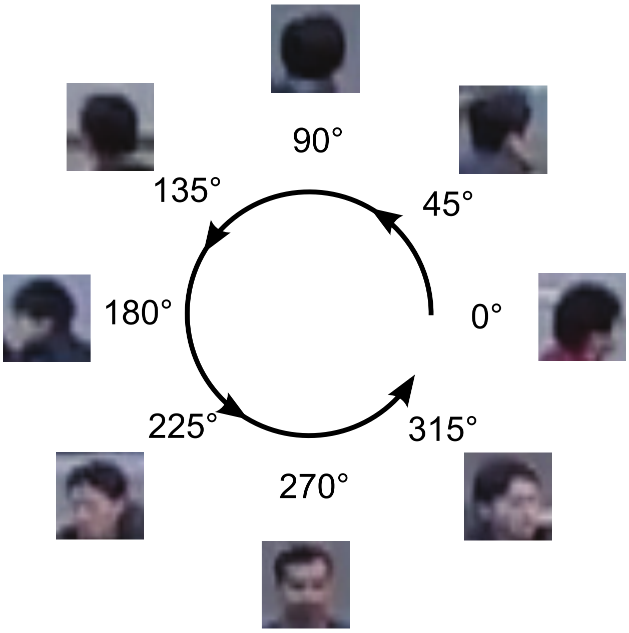


Figure 2.1: Illustration of a head pose regression task in which head pose is defined as continuous angle values in an image plane.

camera positions, head appearances in different scenes and even within the same scene can be significantly different. An example of such differences in appearance of people within the same scene is shown in Figure 2.2. Even though pedestrians are walking in the same direction, their head appearances are different when captured from different locations. In other words, if lighting conditions or camera positions are significantly different between the locations where training and test images are taken, mappings between the direction and the appearance would also be different. Nevertheless, it is not always possible to collect training samples for every test case.

The framework of our proposed method is summarized in Figure 2.3. Our method acquires training data from an input video sequence by using walking directions as a cue to infer head pose. The scene is then segmented into multiple regions with similar head appearances, and a head pose estimator is constructed for each region.

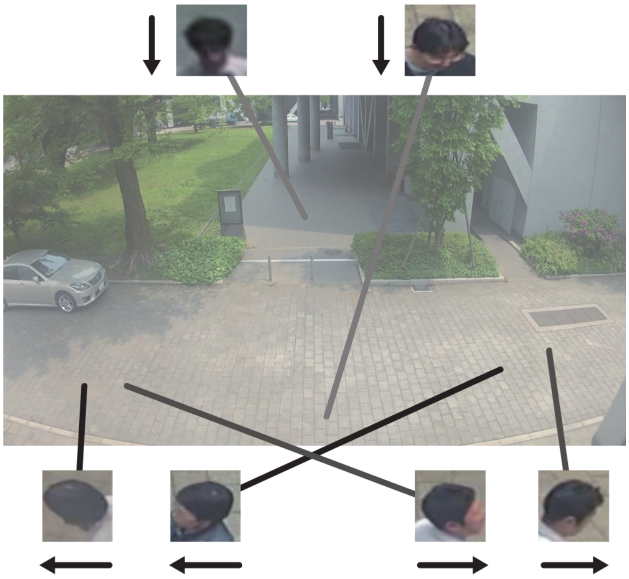


Figure 2.2: An example of differences in appearance of people in a scene. Even when the pedestrians are walking in the same direction, the head appearance is different when captured from different locations in the scene.

In order to obtain walking trajectories of pedestrians in the video, we employed the head tracking method by Benfold and Reid [32]. The method is based on a Kalman filter [58] with two types of measurements: the head locations given by a HOG-based head detector [59] and the velocity of head motion computed from multiple corner features [60, 61]. In each frame, a head image I, a head location $\boldsymbol{u} = (x, y)$ in the image plane, and a measurement error $\boldsymbol{c} = (c^{(x)}, c^{(y)})$ are collected for analysis, where the terms $c^{(x)}$ and $c^{(y)}$ are the respective variances of the measurement on x and y axes of the Kalman filter. The pedestrian tracking algorithm is applied to the entire input sequence, and a trajectory, *i.e.*, a set of head images $\{I_1, \ldots, I_N\}$, head locations $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N\}$ and error measurements $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N\}$, is acquired for each pedestrian. Here, N denotes the length of the trajectory and it varies depending on the trajectory.

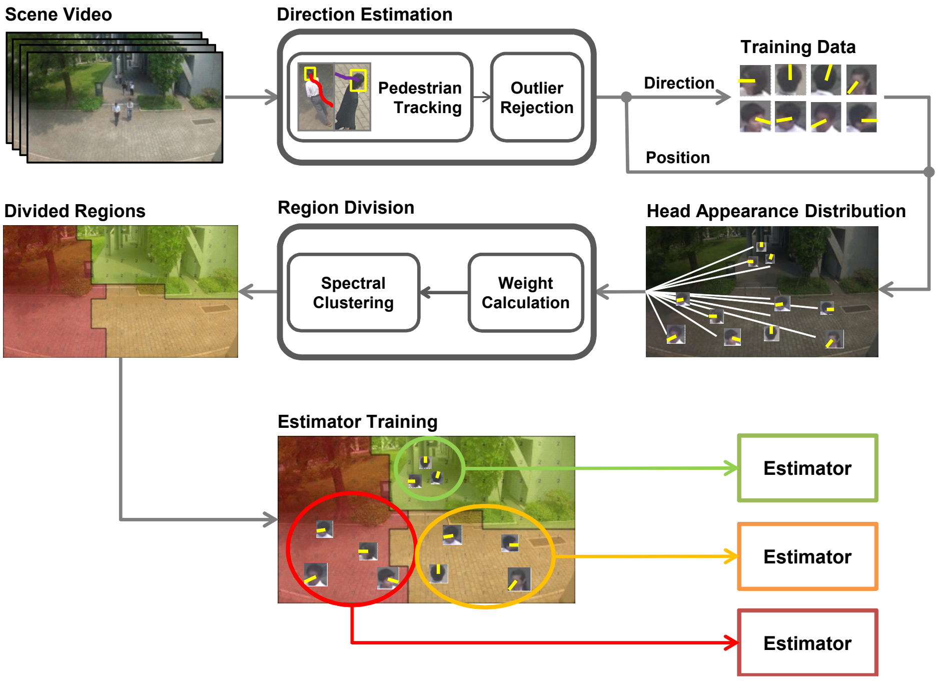


Figure 2.3: The proposed framework. From an input video sequence, our method acquires head pose training data by using walking directions as a cue to infer head pose. The scene is then segmented into multiple regions with similar head appearance on the basis of acquired training samples, and head pose estimators are constructed separately for each region.

2.3 Training Data Acquisition

This section describes our technique to aggregate a scene-specific dataset. Given tracked trajectories of pedestrians, we estimate their walking direction, which can be assumed to indicate their head pose in the images. Erroneous samples that will cause errors in the trained estimators are rejected, and then we collect the remaining training samples to construct the head pose dataset. The proposed method is described in more detail in the following sections.

2.3.1 Estimating Walking Directions

To account for the fact that pedestrians do not always walk straight, our method first divides each possibly curved trajectory into straight line segments. More specifically, each trajectory S is divided into segments $\{S_1, \ldots, S_M\}$ by polyline simplification using the Douglas-Peucker algorithm [62]. This algorithm constructs a minimal set of lines so that the orthogonal distances from each point to its nearest line is less than a given threshold $d_{\max}$. Since pedestrians get to appear smaller as they move away from the camera, the threshold $d_{\max}$ should be defined in a location-dependent way. Therefore, based on the fact that the physical size of the curve is proportional to the observed head size, we define the threshold $d_{\max}$ as $d_{\max} = \tau_p \cdot \bar{s}_t$, where τ_p is a scale-invariant constant and $\bar{s}_t = \sum_{i=1}^{N} \sqrt{s_x(\boldsymbol{u}_i) \cdot s_y(\boldsymbol{u}_i)}/N$ is the average head length calculated assuming a square shape observed over a trajectory. $s_x(\boldsymbol{u})$ and $s_y(\boldsymbol{u})$ are the expected width and height of the head at the position $\boldsymbol{u}$. They are calculated assuming that the average human height is 1.7 meters, and heads are modeled as cylinders that are 22.0 centimeters tall and 20.0 centimeters in diameter, in the same manner as [32]. An example of polyline simplification is shown in Figure 2.4. In the figure, the curved line shows the raw tracking result, and the straight lines show line segments obtained using the polyline simplification algorithm.

2.3.2 Rejecting Outlier Segments

Walking directions obtained from polyline simplification of a trajectory do not always correspond to head orientations since people can move their heads freely even while they are walking. This brings errors in the training labels. Head pose estimation algorithms are not always robust to such outliers, and thus it is preferable to reject them prior to the learning stage. To address this problem, we introduce a strategy to reject unreliable segments from the

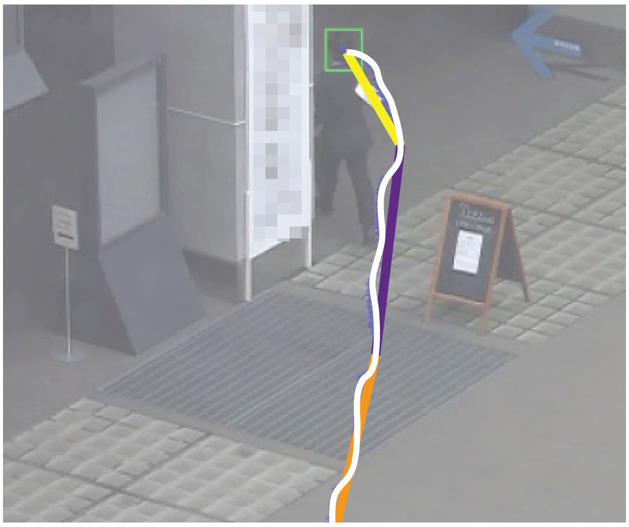


Figure 2.4: An example of polyline simplification. The curved line shows a tracking result and straight lines show the result of polyline simplification.

tracking results.

There are three kinds of segments that cause erroneous training samples: 1) segments with large tracking errors, 2) segments with short length or slow movement, and 3) segments with large image variance. The details of each kind are as follows. Let us assume that a segment contains T head locations $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T\}$.

Segments with large tracking errors: Although the pedestrian trackers can resume their tracking and are robust to a few mis-detections, a large number of mis-detections can produce erroneous trajectories and poor head image localizations. These situations will result in a large number of erroneous points and large line fitting errors, which should be rejected.

To calculate the number of erroneous points, a point $\boldsymbol{u}_t$ is identified to be erroneous if the error measurement of the tracker is significantly large:

$$\sqrt{\left(\frac{c_t^{(x)}}{s_x(\boldsymbol{u}_t)}\right)^2 + \left(\frac{c_t^{(y)}}{s_y(\boldsymbol{u}_t)}\right)^2} > \alpha. \tag{2.1}$$

where $\boldsymbol{c}_t = (c_t^{(x)}, c_t^{(y)})$ is the measurement error of the tracker. Since these measurement errors should be evaluated according to their physical size, head width $s_x(\boldsymbol{u}_t)$ and height $s_y(\boldsymbol{u}_t)$ at the location of the tracker $\boldsymbol{u}_t$ are introduced to scale the measurements.

Using this measure, we reject segments if the ratio of erroneous segment points to the total number of points in the segment is larger than a predefined threshold τ_e. Note that α indicates the acceptable error level, while τ_e controls the number of acceptable erroneous points in the segment. These parameters are not independent to each other, thus we first chose α to reject trajectory points where the head detector failed. When the head detector of the tracker module failed twice in a row, α is set to the error level of the tracker at that moment. After α is selected, τ_e is then chosen accordingly.

To reject segments with large line fitting errors, we calculate the summation of the orthogonal distances from each point to the estimated line over a segment and divide the summation by the length of the segment. We then reject a segment if

$$\frac{1}{|\boldsymbol{u}_T - \boldsymbol{u}_1|} \cdot \sum_{t=1}^{T} \frac{|a \cdot x_t + b \cdot y_t + c|}{\sqrt{a^2 + b^2}} \geq \tau_l, \tag{2.2}$$

where τ_l is a threshold indicating the maximum acceptable level of line fitting errors. $\boldsymbol{u}_t = (x_t, y_t)$ is a point in the segment. The left-hand side of equation (2.2) is a scale-independent line fitting error of the estimated line $ax + by + c = 0$.

Segments with short length or slow movement: Pedestrians making slow or no movements are often seen in a scene, *e.g.*, people talking to each other in the same location. Using the walking direction to estimate head pose in such situations would give erroneous results. Therefore, segments that are short in length or that have slow movements need to be rejected. Rejecting segments with short length also removes cases where false positive objects are detected as heads, which usually stay within a small area. Therefore, we reject a segment if

$$\frac{|\boldsymbol{u}_T - \boldsymbol{u}_1|}{\bar{s}_s} \leq \tau_n \text{ or } \frac{|\boldsymbol{u}_T - \boldsymbol{u}_1|}{T \cdot \bar{s}_s} \leq \tau_v \tag{2.3}$$

where τ_n and τ_v are predefined thresholds for detection of segments with short length and slow movements, respectively, and $\bar{s}_s = \sum_{i=1}^{T} \sqrt{s_x(\boldsymbol{u}_i) \cdot s_y(\boldsymbol{u}_i)}/T$ is the average of the head-length factors over the segment.

Segments with large image variance: Pedestrians in the video are sometimes observed turning their head while they walk, which also leads to erroneous direction estimation results. Because large variations in head appearance are expected in such cases, segments with large image variations should be rejected. We calculate the variance of resized head image vectors $\{\hat{\mathbf{I}}\}$ whose dimensions are denoted by C. A segment is considered to have large variance if

$$\frac{\sum_{t=1}^{T} |\hat{\mathbf{I}}_t - \bar{\mathbf{I}}|^2}{T \cdot C} \geq \tau_{var}, \tag{2.4}$$

where $\hat{\mathbf{I}}_t$ denotes the t-th resized image, $\bar{\mathbf{I}}$ is a mean image calculated from all resized images $\hat{\mathbf{I}}$ in the segment, and τ_{var} is a predefined constant. While high τ_{var} makes the algorithm accept more samples, low τ_{var} makes the algorithm more selective about the stability of head images.

2.3.3 Representative Image Selection

Most outlier segments are rejected in the outlier segment rejection process, and the remaining segments contain correct data. Since only one orientation is assigned to each segment, most of the images in each segment are redundant. Using all the images for training would result in an excessively large dataset, which increases the computational time for many machine learning tools. Therefore, one representative image per segment is selected and used as training data.

In this work, we select the image that is most similar to the mean image of the segment. For each segment, the Mahalanobis distance from the mean image is calculated for every resized image $\hat{\mathbf{I}}_t$ in the segment and the image with the lowest distance is selected. This enables us to select the representative image while avoiding effects that can be seen in the mean image, *e.g.*, blur or distortion.

2.4 Adaptive Scene Segmentation for Localized Direction Estimation

As mentioned before, appearance differences of training samples in the scene can reduce the accuracy of the estimator. This section addresses our approach of segmenting a scene into multiple regions in each of which the heads with the same direction have a similar appearance. Because there is no definitive way to define regions with similar head appearances, an unsupervised clustering approach is taken to segment a scene into such regions. In this work, spectral clustering is used to segment a scene. Given a set of points and a similarity matrix defining the similarity of each pair of points, spectral clustering techniques cluster the set into disjoint subsets with high intra-cluster similarity and low inter-cluster similarity. Normalized cut [63],

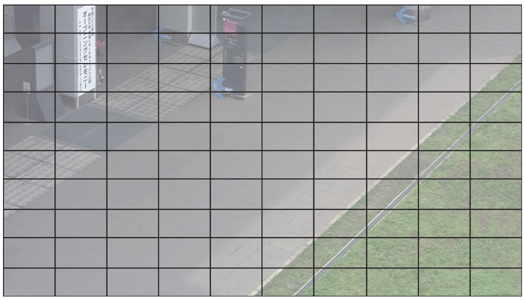


Figure 2.5: An example of dividing a scene into 10×10 unit regions ($K = 100$). These regions serves as the smallest unit to construct each region.

which is one of the most common spectral clustering algorithms, is applied in our approach.

Our approach first divides the scene into K rectangular unit regions $V = \{v_1, \ldots, v_K\}$ which are used as the set of nodes. Figure 2.5 shows an example of 10×10 unit regions ($K = 100$). Then normalized cut is applied to cluster the regions V into R clusters, $\mathbf{A} = \{A_1, \ldots, A_R\}$, where $A_i \neq \emptyset$, $A_i \subset V$, $A_i \cap A_j = \emptyset$ $(1 \leq \forall i, j \leq R, i \neq j)$ and $\bigcup_{i=1}^{R} A_i = V$. In the following sections, we discuss how to calculate the similarity weight function $w(v_i, v_j)$ for each pair of unit regions and how to choose the appropriate number of regions.

2.4.1 Weight Function

With the dataset D obtained using the method described in Section 2.3, we define D_v as training samples captured at the unit region v. Our proposed similarity weight $w(v_i, v_j)$ between two unit regions v_i and v_j is defined with the distance weight w_d and the sample weight w_s as $w(v_i, v_j) = w_d(v_i, v_j) \cdot w_s(v_i, v_j)$.

The distance weight, $w_d(v_i, v_j)$, measures how close two unit regions v_i and v_j are. This takes into account the fact that training samples acquired from nearby locations tend to be more similar than those acquired from distant locations. The distance weight also makes segmented regions spatially smooth. We define distance weight w_d as follows:

$$w_d(v_i, v_j) = e^{-\frac{\|\boldsymbol{X}_i - \boldsymbol{X}_j\|_2^2}{\sigma_d}}, \tag{2.5}$$

where $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are the positions of the unit regions v_i and v_j, respectively and σ_d is a predefined constant.

Sample weight $w_s(v_i, v_j)$ measures the similarity between training samples acquired from the two unit regions v_i and v_j. Two unit regions v_i and v_j should be merged into the same region if the training samples D_{v_i} are similar to D_{v_j}. Sample weight is defined as

$$w_s(v_i, v_j) = e^{-\frac{d_{\mathrm{u}}(v_i, v_j)}{\sigma_s}}, \tag{2.6}$$

where σ_s is a constant and $d_{\mathrm{u}}(v_i, v_j)$ is a function that measures the difference between samples in two unit regions. The comparison is done between training samples corresponding to similar head pose, *i.e.*,

$$d_{\mathrm{u}}(v_i, v_j) = \frac{\sum\limits_{(\boldsymbol{h}_i, p_i) \in D_{v_i}, (\boldsymbol{h}_j, p_j) \in D_{v_j}} d(\boldsymbol{h}_i, \boldsymbol{h}_j) \cdot \phi(p_i, p_j)}{\sum\limits_{(\boldsymbol{h}_i, p_i) \in D_{v_i}, (\boldsymbol{h}_j, p_j) \in D_{v_j}} \phi(p_i, p_j)}, \tag{2.7}$$

where $(\boldsymbol{h}_i, p_i)$ and $(\boldsymbol{h}_j, p_j)$ are the feature vectors and head pose labels for a training sample in the dataset D_{v_i} and D_{v_j}, respectively. Here, $\phi(p_i, p_j)$ is defined using a threshold[2] θ:

$$\phi(p_i, p_j) = \begin{cases} 1 & \text{if } |p_i - p_j| < \theta \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

The value $d(\boldsymbol{h}_i, \boldsymbol{h}_j)$ measures differences between a pair of samples, and is

[2] In our experiments, head samples with differences less than 45.0 degrees were defined as being similar, *i.e.*, we set $\theta = 45.0$ (degrees).

defined as the weighted distance between the feature vectors:

$$d(\boldsymbol{h}_i, \boldsymbol{h}_j) = \sqrt{(\boldsymbol{h}_i - \boldsymbol{h}_j)^T M (\boldsymbol{h}_i - \boldsymbol{h}_j)}, \tag{2.9}$$

where $M = \mathrm{diag}[M_i]$ is the diagonal matrix indicating the importance of each feature in the feature vectors. M_i should be large if the i-th feature has a strong impact on distinguishing head pose. Although the importance matrix M can be obtained by using several approaches, in this work, M was obtained from the variable importance vector calculated from the random trees estimator [64], which was trained using the whole dataset D.

2.4.2 Determining the Number of Regions

In addition to the weight function w, it is also important to select the appropriate number of regions. It is preferable for a scene to be segmented into as many regions as possible to take advantage of having samples with similar appearances inside the same region. However, if a region is too small, the number of training samples will be insufficient, and the trained estimators will have significant generalization errors.

We perform cross validation on the scene segmented with different numbers of regions and select the one that minimizes the cross-validation error. The cross validation errors is defined as the weighted sum of the validations errors in each region: for a segmentation $\boldsymbol{A} = \{A_1, A_2, \ldots, A_R\}$,

$$E_{\mathrm{g}}(R) = \frac{1}{|D|} \sum_{r=1}^{R} E_{\mathrm{c}}(D_r) \cdot |D_r|, \tag{2.10}$$

where D_r is the set of training samples captured within region A_r, and $E_{\mathrm{c}}(D_r)$ is the 5-fold cross validation error using the training data D_r. For each sequence, cross-validation errors for scene segmentation with R ($1 \leq R \leq R_{\max}$) are calculated, and the number R^* that minimizes the cross-validation error is then selected, *i.e.*, $R^* = \arg\min_R E_{\rm g}(R)$. We consider head poses estimated by using our proposed method in Section 2.3 as ground-truth data, and therefore we do not use manually-labeled ground truth data for computing cross-validation errors.

2.4.3 Training Estimators

As a result of the above processes, we obtain a set of regions $\boldsymbol{A}' = \{A'_1, \ldots, A'_{R^*}\}$ in each of which the appearance of the training samples is similar. Estimators $f_1, \ldots, f_{R^*}$ are then created for each region, and each of them is trained with the samples in its corresponding region; *i.e.*, estimator f_j for region A'_j is trained with the dataset $D_{train,j} = \bigcup\{D_v; v \in A'_j\}$. The estimator in each region is applied for test samples captured in its corresponding region; *i.e.*, the test samples in region A'_j are tested with the estimator f_j. Note that test samples are separated from training samples and are not included in the dataset D.

2.5 Experimental Results

We conducted experiments using five video sequences that were recorded using different cameras in different scenes. The details of each sequence and the numbers of samples obtained as a result of the training data acquisition approach are summarized in Table 2.1. Example frames in the videos are shown in Figure 2.6. Examples of the obtained head images are also shown with the estimated walking direction shown next to the image.

The parameters were set as follows for every scene; $\tau_p = 0.8$, $\alpha = 0.5$, $\tau_l = 5$, $\tau_e = 0.4$, $\sigma_d = 1000$, $\sigma_s = 0.1$, $\tau_n = 3.0$, $\tau_v = 0.02$, $\tau_{var} = 0.0035$. The effect of applying the rejection methods is analyzed in Section 2.5.3, and the robustness against the parameter setting is analyzed in Section 2.5.4.

[3] The Town Centre Sequence was the publicly available sequence from [38].

Table 2.1: Details of sequences used in our experiments. The first three columns show the name, resolution and length of each sequence, respectively. The fourth column indicates the number of test samples captured from each sequence, and the last column shows the number of training samples acquired with the proposed method.

Sequence Name	**Resolution**	**Length (minutes)**	**Test Samples**	**Obtained Samples**
Sequence 1	1920 × 1080	420	300	7841
Sequence 2	1120 × 780	10	100	1312
Sequence 3	1280 × 720	10	135	693
Sequence 4	1920 × 1080	90	305	3075
Town Centre[3]	1920 × 1080	22	4347	6190

2.5.1 Estimation of head pose

To evaluate the performance of our proposed method, we compared our method with regressors trained with a generic dataset that consists of head images collected from other scenes. We constructed a generic dataset using 1477 training samples taken from the Gaze Direction Dataset [65], which was used in [32]. Figure 2.7 shows examples of head images included in the generic dataset.

All of the head images were resized to 40 × 40 pixels ($C = 1600$), and all of the scenes were divided into 16 × 9 unit regions ($K = 144$). An image descriptor similar to the one in [38] was used here. The descriptor is the concatenation of two features. The first feature measures color difference between two pixels at two different locations. The second feature measures difference between two different bins from Histograms of Gradients (HOGs) features extracted from the head images divided into 4 × 4 cell grids and normalized spatially across 2 × 2 blocks of cells. In our work, 400 pairs of points were chosen randomly for each feature.

The experiment was conducted using two regressors: Support Vector Re-

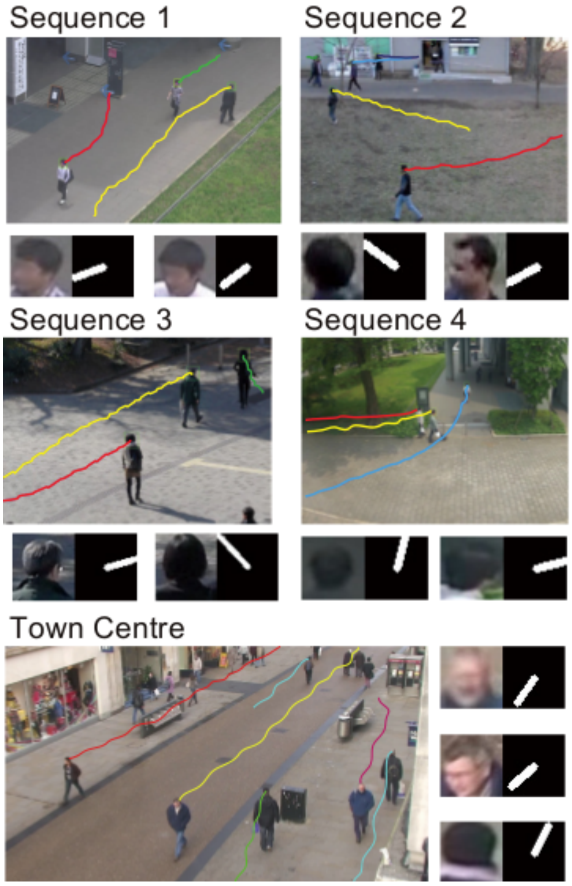


Figure 2.6: Example frames in test video sequences. The input video frames are overlaid with pedestrian tracking results. Examples of obtained head images are also shown; the white line in the right part of each image represents the estimated walking direction.

gression (SVR) and regression with random trees [64]. Both of them were implemented by using OpenCV library [66]. SVR is one of the most common machine learning tools used in head pose and gaze estimation [31, 67]. The combination of random trees and the above-discussed descriptor is similar to the estimator in [38] and was used for a fair comparison between our results with those reported in [38]. Our method finished training within 10 minutes, and testing took less than 1 ms per test sample on an Intel Core 2 Duo 3.00GHz CPU.

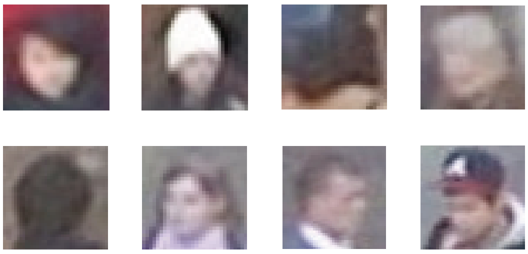


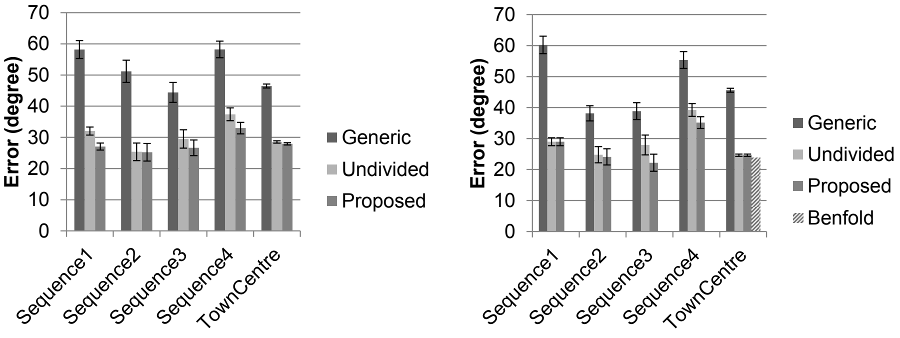
Figure 2.7: Example images in the Gaze Direction dataset.

A comparison of the mean absolute angle errors (MAAE) between our method and the baseline using the generic dataset is summarized in Figure 2.8. Here, we also compared the results without using the scene segmentation method. The **Generic** results were calculated based on regressors trained using the generic dataset, the **Undivided** results were calculated based on regressors trained using samples acquired without scene segmentation, and **Proposed** results were calculated based on our proposed method. **Benfold** result shows the angle error stated in [38].

The graphs show that the errors in regression tasks using the dataset obtained with our method are significantly smaller than those of the generic dataset. Scene segmentation further reduces errors for scenes with large variations in lighting conditions, such as sequence 3, or large differences in camera viewing angles such as sequence 4. Our result is comparable to that of Benfold and Reid [38].

2.5.2 Adaptive Scene Segmentation

To test the effectiveness of region segmentation in our method, we measured the relation between cross-validation errors and actual estimation errors. We applied our method with different numbers of regions and recorded their respective cross validation errors. In our experiments, we set the maximum number of regions to calculate cross-validation errors to 10. The com-

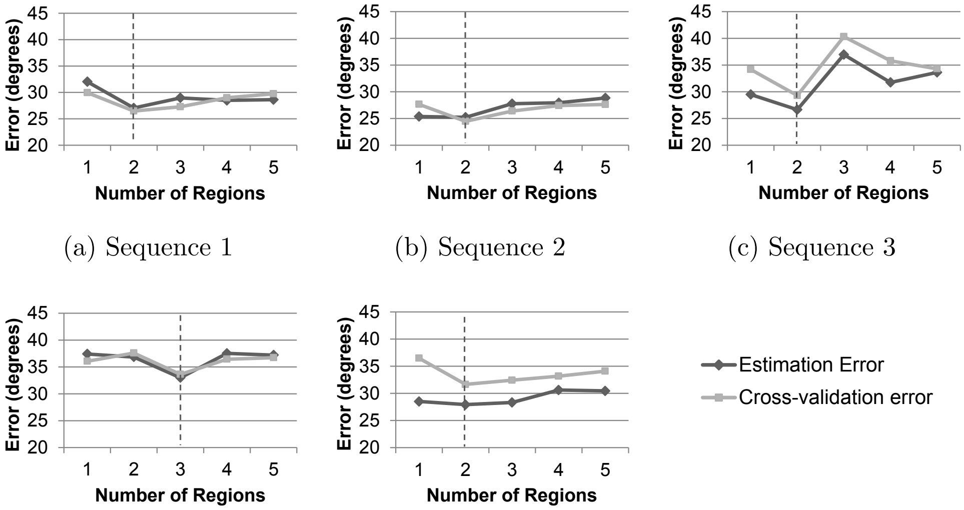


(a) SVR regressor

(b) Random trees regressor

Figure 2.8: Errors of head pose estimation for SVR and random trees regressors. **Generic** results were calculated based on regressors trained using the generic dataset; **Undivided** results were calculated based on our sample collection approach without scene segmentation; **Proposed** results were calculated based on regressors trained using samples acquired with our method, and **Benfold** result shows the angle error stated by Benfold and Reid [38] on Town Centre dataset using their proposed method. The errors were measured using the mean absolute angle error (MAAE). Standard errors are indicated as error bars.

parison of estimation errors and cross-validation errors for SVR with different number of regions are shown in Figure 2.9. Both the cross-validation errors and the estimation errors increase when the number of regions increases to more than 5 and have been omitted from the graph for clearer representation. The number of regions that minimizes cross validation errors was chosen as the optimum number of regions. It can be seen that minimizing the cross-validation errors on training samples also minimize the estimation errors on test samples. The results of scene segmentation are shown in Figure 2.10. It can be seen that in sequences 3 and 4, areas with a large camera angle or illumination differences were segmented automatically. No significant change in performance was seen in the other sequences where head appearances remain relatively uniform in the scene.



(a) Sequence 1

(b) Sequence 2

(c) Sequence 3

(d) Sequence 4

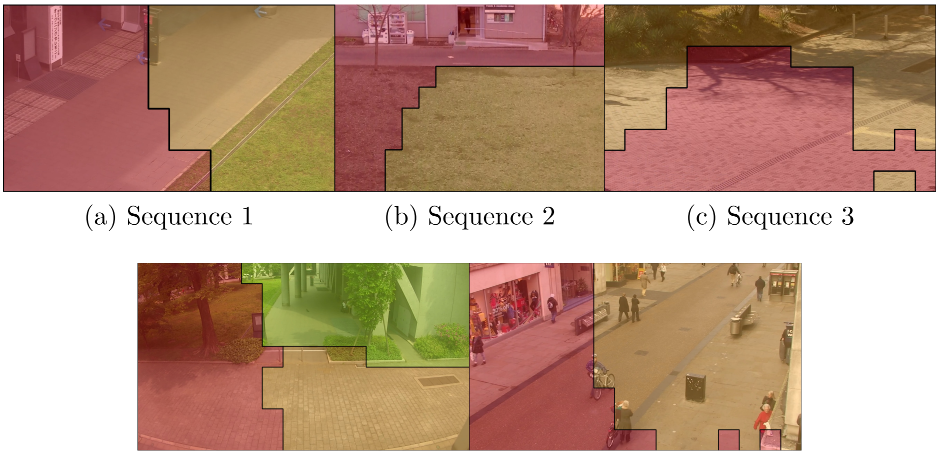
(e) Town Centre

Figure 2.9: Comparison of estimation errors and cross-validation errors with varying number of regions. The values of R^* where cross validation errors were minimum were selected and are shown as dotted lines in the graphs.

2.5.3 Analysis of Outlier Rejection

To measure the effectiveness of our outlier rejection rules descibed in Section 2.3.2, we tested our proposed method with omitting each rule. An example result from Sequence 4 is shown in Figure 2.11. Similar trends were observed in the other datasets, although we did not include those results here. It can be seen that our proposed method achieves the best estimation accuracy while maintaining the smallest dataset size.

These results indicate that short segment length and slow movement criteria significantly reduce estimation errors. This is intuitively reasonable because these rules reject trajectories where pedestrians are talking to each other, which are often observed in scenes. Rejection of short length segments also further reduces the errors by rejecting trajectories generated by false positive objects. In addition, it can be seen that rejection of samples with



(a) Sequence 1 (b) Sequence 2 (c) Sequence 3

(d) Sequence 4 (e) Town Centre

Figure 2.10: Scene segmentation results with the proposed method. The figure is best viewed in color.

high variance significantly reduces the number of captured samples. This improves the training speed for large datasets.

2.5.4 Analysis of Parameter Settings

In this section, the effects of different parameter values on the results are analyzed. We conducted experiments by applying the proposed method and changing each parameter value by twenty percent. We did not perform the analysis on the Town Centre dataset because the tracking results provided by the authors were used, and tracker error variance values were not available.

We show two example results of parameter tests in Figure 2.12. In the figure, the center columns with dotted lines show the default value mentioned in Section 2.5. The left and right columns show the default values that were changed by twenty percents. In each experiment, only one parameter was changed, and the other parameters were kept at their default values. Generally speaking, if the parameters are set too strictly, estimation errors increase due to the lack of sample variations. If the parameters are set to be

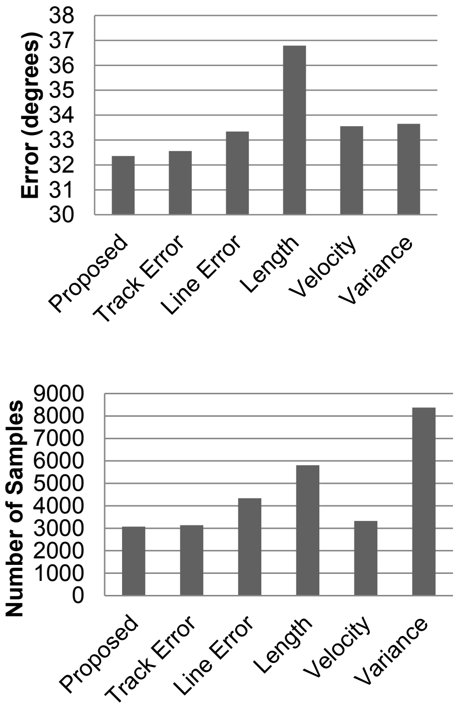
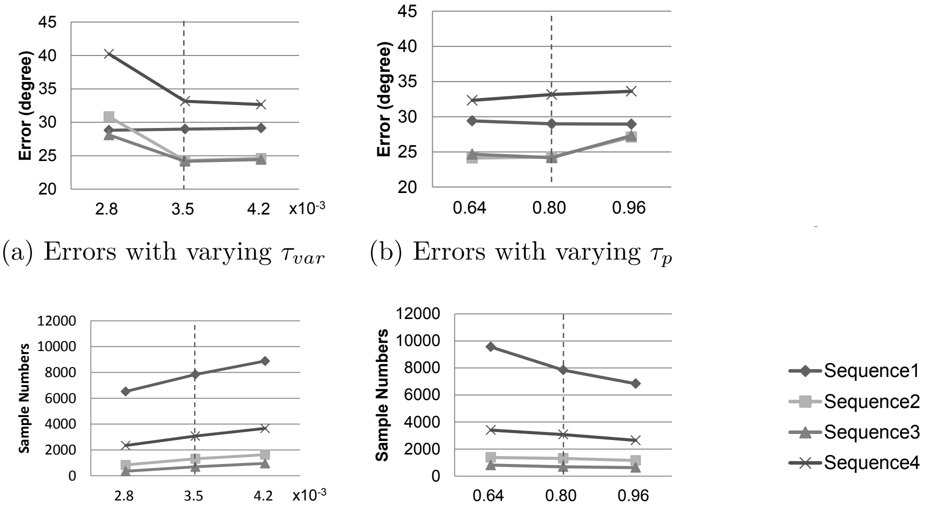


Figure 2.11: Errors and number of samples captured from Sequence 4 with various outlier rejection methods omitted using SVR as regressor. **Proposed** column shows our proposed method. **Track Error** and **Line Error** columns omitted the rejection of segments with erroneous tracking and large line fitting errors, respectively. **Length** and **Velocity** columns omitted the rejection of segments with short length and slow movement, respectively. **Variance** column omitted the rejection of segments with high image variance.

more tolerant, the number of samples increases while the estimation errors are not significantly reduced.

Although rejecting segments with large variance reduces estimation errors as stated in Section 2.5.3, it is apparent that estimation errors significantly increase with a stricter threshold. This indicates that although image variance helps in rejecting images with incorrect head pose, variations in head appearance are also important for training regressors. Increasing the threshold value by twenty percent, however, did not significantly affect the estimation result.



(a) Errors with varying τ_{var} (b) Errors with varying τ_p

(c) Number of samples with varying τ_{var} (d) Number of samples with varying τ_p

Figure 2.12: Variations of estimation errors and number of collected samples with varying parameters. Center columns with dotted lines indicate the default value of each parameter.

This is because image segments where people turn their head usually have large variance, and thus, there is a large margin for the variance threshold to reject such segments. Increasing the polyline fitting threshold will cause curved lines to be estimated as straight lines. This significantly increased estimation errors, especially in sequence 2 and sequence 3 which contain a small number of samples. Reducing the threshold increased the number of samples but did not significantly reduce estimation errors. This is because if the line estimated by polyline simplification is sufficiently straight, reducing the threshold will further divide the line but will not yield any benefits.

2.6 Conclusion

We proposed a method of appearance-based head pose estimation that can automatically adapt to test scenes. The key idea behind the proposed framework is to use walking directions as a cue to infer head pose. A pedestrian tracker is first applied to the input video sequence, and then head pose for each pedestrian is estimated based on his/her walking direction. Outlier segments are rejected, and then a scene-specific dataset of head images labeled by their walking directions is automatically acquired. Each scene is then segmented into multiple regions according to the appearance of acquired head images with the same direction. Finally, a head pose estimator for each region is created by using training samples acquired from that region. The results of our experiments verified that our method estimates head pose accurately without any need to manually collect a ground-truth dataset in real scenes. This is a significant advantage compared to existing methods when applied to practical scenarios.

Appearance-based head pose estimation from low-resolution images is itself a difficult task, and there is still room for improvement in both feature description and estimation techniques. We believe that investigating the learning algorithm itself is an important future task.

Chapter 3

Social Group Discovery using Attention-based Cues

3.1 Introduction

Group discovery has recently become an active research topic in computer vision literatures. In order to understand how humans form social groups, it is important to distinguish behaviors of humans in the same social group from behaviors of those in a different social group. It can be observed that humans in the same social group tend to act in unison rather than individually and activities performed in social groups also differ from those performed by individuals. For example, humans from the same social group tend to talk and pay attention to one another, while ignoring those from different social groups. They also tend to stay in close proximity and their motion direction tends to be similar. These social group behaviors suggest that social group discovery can be performed with two types of visual cues, attention-based cues, such as how pedestrians pay attention to one another, and position-based cues, such as relative distances or movement direction between pedestrians.

Recent approaches have demonstrated that position-based cues can be applied to the social group discovery problem [8, 7, 12, 13, 14, 15]. While position-based cues are useful for discovering social groups in scenes with distinctive group movements, they do not work well in scenes where social groups in close proximity are standing still or have similar movement directions. Social events also tend to cause unstable trajectory of pedestrians in social groups, and make them difficult to be discovered using position-based cues alone. Figure 3.1 has 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. This makes it hard to robustly estimate the social group using position-based cues alone. However, social behaviors such as a conversation between pedestrians can be a great cue to discovering this social group.

Recent advances have shown that social behaviors such as group conversations or group discussions can be effectively estimated by using attention-based cues [19, 20]. This suggests their great potential to solve the problem of social group discovery. Therefore, we propose that attention-based cues be incorporated into our approach to social group discovery. Several researchers in the computer vision literature have estimated human attention from their gazes [22, 23, 24]. However, stable gaze estimates are difficult to be obtained from surveillance videos where high-level features such as eye positions cannot be accurately obtained, and head pose estimates of humans are often used instead of their gazes in order to approximate human attention in such videos [25, 26, 27, 28]. Appearance-based approaches to estimating head poses have demonstrated significant advantages over other approaches in low resolution images [50], and recent advances have allowed us to robustly infer the head poses of pedestrians in real time even without having to use manually prepared training data [38, 39, 40]. We therefore employed the appearance-based approach to acquire the head poses of pedestrians in

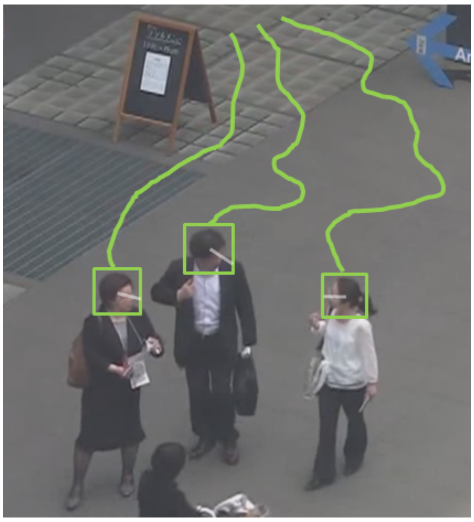


Figure 3.1: Example of ambiguous group relationship without head pose information. Squares and curved lines indicate tracked trajectories of pedestrians in social group. Straight lines in squares indicate estimated head poses. Relative distance between rightmost person and rest varies greatly over trajectories. This makes it hard to robustly estimate this social group using position-based cues alone. Attention-based cues such as people's eye gazes strongly suggest social groups, and can be used to help in this case.

videos.

After attention-based and position-based cues have been acquired, our approach combines them to learn human behavior models and discover social groups in videos. This is the first work, to the best of our knowledge, to propose: 1) a method that uses the 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 social group discovery task. We used a set of basic measurements obtained from the cues, and implicitly trained a set of decision trees by using a supervised learning algorithm without enforcing explicit modeling of social group behaviors.

Our proposed approach was extended from that by Chamveha *et al.* [68] with the following additions: 1) The previous work [68] focused on discovering pairwise social group relationships between pairs of pedestrians, but did not address how to find all of the members of social groups. Our extension addresses this problem by modeling social relationships of pedestrians using a graph representation and perform a clustering algorithm to discover social groups and all of their members in a video. We also measured the results with different variants of the clustering algorithm. 2) The effectiveness of each attention-based cue is more carefully studied. 3) Moreover, our proposed approach is more thoroughly tested on the UCLA Courtyard dataset [69], in addition to the two datasets used by Chamveha *et al.* [68].

3.2 Proposed Framework

The framework for our method is outlined in Figure 3.2. We first collect measurements based on two types of cues from the training video: *attention-based cues* and *position-based cues*. Attention-based cues are derived from human behaviors related to their attention, and position-based cues are derived from the observed trajectories of pedestrians. We then aggregate these measurements into feature vectors which are then used learn the group behavior models. We calculate the group probability scores for every pair of pedestrians in the test scene by using the constructed models, and clustered the pedestrians into groups using a graph clustering approach.

With $\boldsymbol{Q} = \{q_1, q_2, \ldots, q_n\}$ defined as the set of pedestrians in the video, we define a social group as $G_i = \{q_{i1}, q_{i2}, \ldots, q_{im}\}$, where $q_{ij} \in \boldsymbol{Q}$, $G_i \in \boldsymbol{G}$, where $\boldsymbol{G} = \{G_1, G_2, \ldots, G_N\}$ is a set of all social groups in the video. We define group mapping function $g : \boldsymbol{Q} \to \boldsymbol{G}$ as a map from the pedestrians to their respective group, *i.e.*, $g(q) = G \iff q \in G$. In this work, we determine social groups given information on past states of pedestrians. Specifically, given past states $\{s_t\}$ of pedestrians in the video, the goal is to find the mapping function $\hat{g}$ that estimates the real mapping function g for pedestrians in the video.

To discover mapping $\hat{g}$, we model social group relationships among pedestrians as a weighted graph $K = (V, E)$ with its corresponding weights W, which we will later refer to as *group probability scores*. Graph nodes $V = \{v_1, v_2, \ldots, v_n\}$ represent pedestrians in the video, where each node v_i represents a pedestrian q_i. Edges E represent relationships between pedestrians. Each edge $e_{ij} = (v_i, v_j)$ connecting two nodes is associated with its group probability score $w_{ij} \in W$ specifying the likelihood that their corresponding pedestrians q_i and q_j are from the same group. Note that edges $e_{ij} = (v_i, v_j)$ only exist if pedestrians q_i and q_j share a nonzero existence, *i.e.*, they simultaneously appear in the video for at least one frame.

Edges connecting pedestrians from the same group will have a group probability score of one in the optimal case, while edges connecting pedestrians from different groups will have a group probability score of zero, *i.e.*,

$$w_{ij} = \begin{cases} 1, & g(v_i) = g(v_j) \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

Group probability scores are crucial to the accuracy of the algorithm for social group discovery and therefore it is important to have precise group probability scores for each pair of pedestrians. However, it is almost impossible in reality to obtain group probability scores explicitly for every pair of pedestrians. Therefore, we propose the group probability score w_{ij} for each edge e_{ij} be inferred from the given pair of pedestrian states $\{s_t^{(i)}\}$ and $\{s_t^{(j)}\}$ of pedestrians q_i and q_j. The scores are calculated based on the attention-based and position-based cues of both pedestrians. Group mapping function $\hat{g}$ is then recovered through a graph clustering approach.

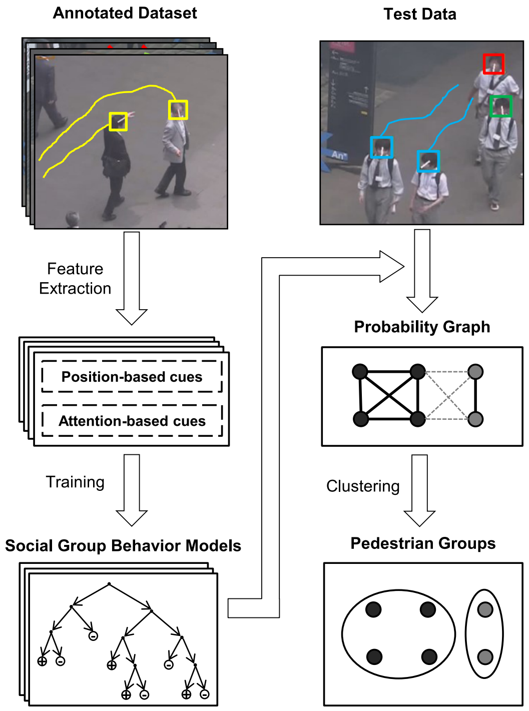


Figure 3.2: Proposed framework. Pedestrian behavior model is first constructed from two types of cues: *attention-based cues* and *position-based cues*. Attention-based cues are derived from observed human behaviors related to attention, and position-based cues are derived from observed trajectories of pedestrians. Group probability scores are calculated for every pair of pedestrians in scene by using constructed model, and pedestrians are clustered into groups with graph clustering approach.

3.2.1 Group Clustering

This section describes the approach to clustering pedestrians into social groups based on their relative group probability scores. The core idea underlying the clustering approach is that pairs of pedestrians with high group probability scores should be assigned to the same social group, while those with low scores should be assigned to different social groups. However, because pedestrians in the same social group do not always have strong social behaviors toward the others, high group probability scores cannot be ex-

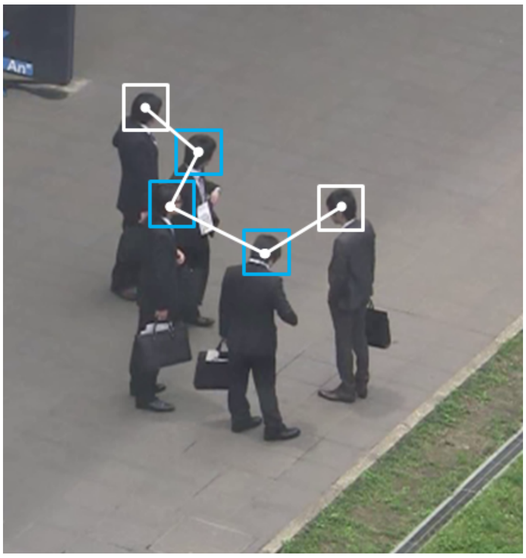


Figure 3.3: Example of social group where pedestrians do not always have strong social behavior toward others. White lines indicate pairs of pedestrians with high group probability scores. Although pedestrians in white squares do not have high group probability scores, they are correctly assigned same social group by transitional property of social group.

pected from every pair of pedestrians in the same social group. Figure 3.3 has an example of such a case, where the white lines indicate pairs of pedestrians with high group probability scores. It can be seen that the pair of pedestrians in white squares do not have strong social behaviors toward each other and it is therefore difficult to evaluate whether these two belong to the same social group or not. However, it can be seen that the relationship between the two pedestrians can be discovered through the chain of relationships of the others in the social group. This chain can be seen as a transitional property of the social group, *i.e.*, if pedestrians q_1 and q_2 are in the same social group and q_2 and q_3 are in the same social group, q_1 and q_3 should also be in the same social group.

We use agglomerative hierarchical clustering algorithm with a generalized version of the single-linkage criteria [70] to address the transitional property of the social group. The hierarchical clustering algorithm is well-known for its flexibility in terms of the freedom it allows in choosing the linkage criteria for merging clusters, and the single-linkage criteria can be satisfactorily generalized to fully address the transitional property of social groups.

With a weighted graph as an input, the algorithm assigns each node in the graph to be in a cluster with itself as the only member. Scores between each pair of clusters, indicating how similar the two clusters are, are then calculated according to the linkage criteria, and the pair with the highest score is then merged. The algorithm repeats the procedure until the highest score is less than a threshold or until every cluster is merged into a single cluster.

In hierarchical clustering with the single-linkage criteria, the score between two clusters is calculated as the maximum group probability scores between pairs of nodes in the cluster. However, hierarchical clustering with the single-linkage criteria is prone to errors, where a single incorrectly estimated weight could cause two different groups to be mistakenly merged into one. Therefore, we generalized the single-linkage criteria to top-k linkage criteria, where the score between each pair of clusters is defined by the average value of k-maximum of the weight between pairs of nodes in the cluster. The single-linkage criteria can be seen as specialized cases of top-k linkage criteria with $k = 1$. We achieve the average-linkage criteria with $k = \infty$, which calculates the average weight between each pair of nodes in the cluster as its score. In other words, we define the score with top-k linkage criteria as $top\_k\_score(G_i, G_j) = \frac{1}{T_k} \sum_{w \in H_{ij}} w$, where T_k is a constant and H_{ij} is the set of T_k highest group probability scores between each pair of nodes in the clusters G_i and G_j. The pseudo-code for the clustering algorithm is given in Algorithm 1.

Algorithm 1 Social group clustering with single-linkage criteria

```
Input: Graph $K = (V, E)$, group probability scores $W$, a constant $T_k$ and
  a clustering threshold $T_w$.
Output: Group assignments $\boldsymbol{G} = \{G_1, G_2, \dots, G_N\}$
  $\boldsymbol{G} := \{G_1, G_2, \dots, G_{|V|}\}$, where $G_i = \{v_i\}$
  while $|\boldsymbol{G}| > 1$ do
    $G_i, G_j := \operatorname*{argmax}_{G_i, G_j \in \boldsymbol{G}}(top\_k\_score(G_i, G_j))$
    $w_{max} := \max(top\_k\_score(G_i, G_j))$
    if $w_{max} < T_w$ then
      return $\boldsymbol{G}$
    else
      Remove $G_i$ and $G_j$ from $\boldsymbol{G}$
      Add $G_i \cup G_j$ to $\boldsymbol{G}$
    end if
  end while
```

3.2.2 Group Probability Score Calculations

Group probability scores between each pair of pedestrians denote the likelihood that they belong to the same social group. It is crucial to precisely calculate such scores to ensure estimates are accurate in the approach. In this approach, the likelihood of two pedestrians being in the same social group are estimated from the behaviors observed from their past states *e.g.*, two pedestrians talking and walking alongside each other for a long period are likely to be in the same social group. Histograms of measurements of attention-based and position-based cues are collected over the entire trajectory to capture these behaviors. The collected histograms are then concatenated as a feature vector to train the estimator for the group probability scores.

Specifically, group probability score for nodes v_i and v_j is calculated from their corresponding 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 q_i and q_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 social group discovery, we model social group behaviors as decision trees so that only informative histogram bins are used for decisions. A random forest regressor [64] was used in our approach to construct the trees. At each time step t, state variable $s_t^{(i)}$ representing pedestrian q_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 q_i, as illustrated in Figure 3.4. We denote the image-plane angle of $\boldsymbol{h}_t^{(i)}$ as $\theta_t^{(i)}$, that of $\boldsymbol{v}_t^{(i)}$ as $\psi_t^{(i)}$, and that of the displacement vector $\boldsymbol{d}_t^{(i,j)} = \boldsymbol{x}_t^{(j)} - \boldsymbol{x}_t^{(i)}$ as $\phi_t^{(i,j)}$. The angles are measured in radians.

3.2.3 Attention-Based Cues

Two types of attention-based cues are exploited in this work. The first cue is the *gaze exchange* between pedestrians. Pedestrians in the same social group often exchange gazes and fix their attention on one another when they are engaged in group events, *e.g.*, conversation events. The second cue is the *mutual attention* of pedestrians in the same social group. This is based on the observation that pedestrians often pay attention to the same object of interest. We took an approach to learning the decision rules of these cues in a supervised manner by using histograms of several measurements, as was discussed earlier. This subsection introduces the details on the measurements, *i.e.*, the required building blocks to model these attention-based cues. For clarity, we have omitted the subscript t from what follows.

Difference between 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 q_i is directly looking at pedestrian q_j, which strongly indicates group events

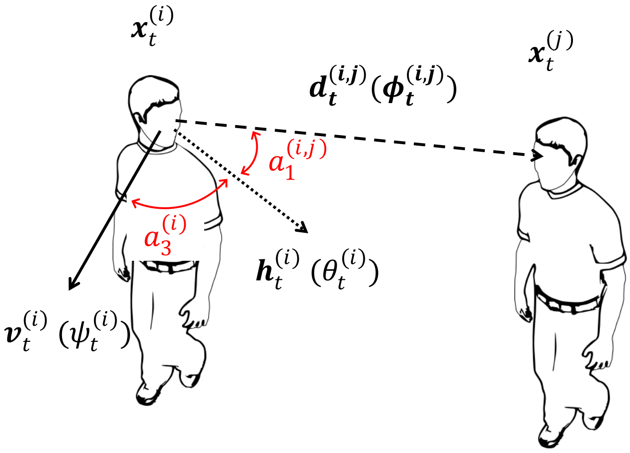

$\boldsymbol{x}_t^{(i)}$
$\boldsymbol{x}_t^{(j)}$
$\boldsymbol{d}_t^{(i,j)}(\boldsymbol{\phi}_t^{(i,j)})$
$a_1^{(i,j)}$
$a_3^{(i)}$
$\boldsymbol{h}_t^{(i)}\ (\theta_t^{(i)})$
$\boldsymbol{v}_t^{(i)}\ (\psi_t^{(i)})$

Figure 3.4: 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. Parentheses next to vectors indicate their corresponding one-dimensional angles. Measurements of $a_1^{(i,j)}$ and $a_3^{(i)}$ have also been provided.

such as a group conversation or a group discussion. This measurement is illustrated in Figure 3.4.

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

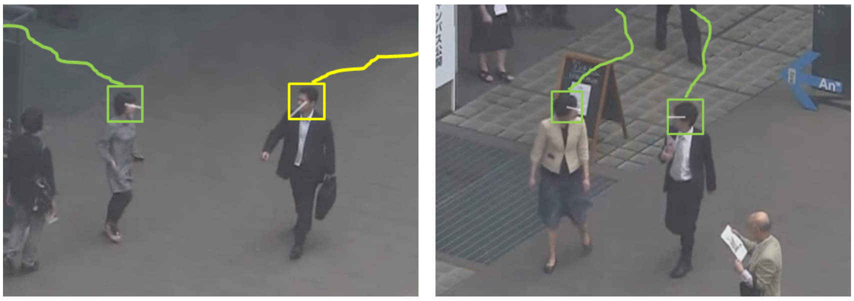
Difference between head pose and walking direction. While the previous two measurements are expected to capture attention-based cues, several different measures are required to obtain efficient decision models. For example, if pedestrians are walking, they naturally tend to look toward the direction in which they are walking. Therefore, looking toward the direction in which they are walking does not suggest that pedestrians are focusing their attention on particular objects or other people in that direction, and the previous measurements are not necessarily informative. We introduced a third measurement to infer the *walking focus* of the pedestrians. This measurement is defined as $a_3^{(i)} = |\psi^{(i)} - \theta^{(i)}|$. This measures how steep pedestrian q_i turns his head away from the direction he/she is walking in. This measurement is illustrated in Figure 3.4.

Figure 3.5a 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 as to whether they are in the same social group. Figure 3.5b 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 social 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 has been included to control the walking focus measurements.

3.2.4 Position-Based Cues

Measurements of position-based cues are derived from the trajectories of two pedestrians as people usually walk in the same direction and at the same speed as people in their group. Recent work by Yamaguchi *et al.* [7] has proposed a set of measurements of position-based cues and demonstrated that it could effectively be used to discover social groups. Therefore, we defined the measurements similarly as follows.



(a) (b)

Figure 3.5: Examples of a_3 measurements. (a) Example when a_1 and a_3 measurements are low for both pedestrians. (b) Example when a_1 measurement is low and a_3 measurement is high for both pedestrians.

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

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

- **Difference between walking 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 social 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 q_i and q_j appear on the scene up to time t, where

Table 3.1: 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 (min)**	**No. of trajectories**	**No. of groups**	**Average group size**
UT-Surveillance	1920 × 1080	75	430	230	1.87
Town Centre	1920 × 1080	22	276	251	1.10
UCLA Courtyard	2560 × 1920	14	125	51	2.45

$T^{(i)} = \{t' | t' \leq t, s_{t'}^{(i)} \neq \emptyset\}$ is a set of time steps where pedestrians q_i appear on the scene up to time t. Pedestrians in the same social group tend to simultaneously enter the scene.

Measurements of the gaze exchange cue a_1, mutual attention cue a_2, walking focus cue a_3, walking speed cue a_4, displacement cue p_1, difference in velocity cue p_2, difference in walking direction cue p_3, difference between walking direction and relative position cue p_4, and time overlap cue p_5 are collected at each time step.

3.2.5 Modeling of Social Behaviors

Measurements for each pair of pedestrians are aggregated into a feature vector in the next step to train the random forest. We want to construct the random forest in such a way that the decision rule of each tree node is based on a single threshold on a measurement, *e.g.*, how often people look at one another with less than τ_h degree angles, and we therefore aggregate each measurement of attention-based cues into a cumulative histogram.

We calculate a set of measurements for every pair of pedestrians with the overlapping existent, *i.e.*, $\{(s_t^{(i)}, s_t^{(j)}) | t \in T_t^{(i,j)}, T_t^{(i,j)} \neq \emptyset\}$ by using an annotated dataset of pedestrian states,. As a_1, a_3, and a_4 measurements are made for each pedestrian, two sets of their measurements are calculated. Therefore, a total number of seven 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 at each time step for measurements of attention-based cues. Each measurement of position-based cues is calculated once and a total five measurements are collected for each pair of pedestrians at each time step.

Each cumulative histogram is constructed with B_a equally-spaced bins. The bins for a_1, a_2, and a_3 are placed between the range $[0, \pi]$. We calculate the maximum speed, v_{max}, for pedestrians in the training set for the a_4 measurement, and the histogram bins for a_4 are placed between the range $[0, v_{max}]$. Measurements of position-based cues are aggregated into standard histograms in the same manner as that by Yamaguchi *et al.* [7]. 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 frames in the video. Histogram bins are placed between the range $[0, 2 \cdot v_{max}]$ for the p_2 measurement, $[0, \pi]$ for the p_3 and p_4 measurements, and $[0, 1]$ for the p_5 measurement. Because training samples contain a different number of frames, both the standard and cumulative histograms are normalized so that the total count in each histogram is summed to 1. The histograms are then aggregated into feature vectors to train the random forest. Feature vectors aggregated from pairs of pedestrians from the same social group are assigned label 1, while those aggregated from pairs of pedestrians from different social group are assigned the label 0.

3.2.6 Pedestrian Tracking and Head Pose Estimates

This section describes how the tracked trajectories as well as head poses of pedestrians were acquired. We chose an approach to pedestrian tracking that was able to provide stable head images, as they are crucial for head pose estimation approaches in crowded scenes. An approach by Benfold and Reid [32] was able to achieve stable tracking results and therefore we employed their approach to generating pedestrian tracks in a video. Their head tracking method was based on a Kalman filter [58] with two types of measurements: the head locations given by a HOG-based head detector [59] and the velocity of head motion computed from multiple corner features [60, 61].

After we had obtained the pedestrian trajectories along with their head images, we applied the unsupervised approach proposed by Chamveha *et al.* [71] to obtain head poses. Their approach automatically aggregated labeled head images by inferring head pose 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 head pose estimator in our approach. Head poses on 2-D image plane were used to approximate of actual head poses in our approach similarly to what had been done earlier [40, 38, 39].

3.3 Experimental Results

We conducted experiments by using three sequences: the UT-Surveillance sequence used by Chamveha *et al.* [71], the Town Centre sequence used by Benfold and Reid [38], and the UCLA Courtyard dataset [69]. The UT-Surveillance sequence contained pedestrians walking along a pathway, often in large groups. The Town Centre sequence contained pedestrians walking along a street. The majority of pedestrians in this dataset walked individually and the dataset therefore contained many negative samples, *i.e.*, pairs of pedestrians that did not belong to the same social group. The UCLA Courtyard dataset contained pedestrians who were engaged in several activ-

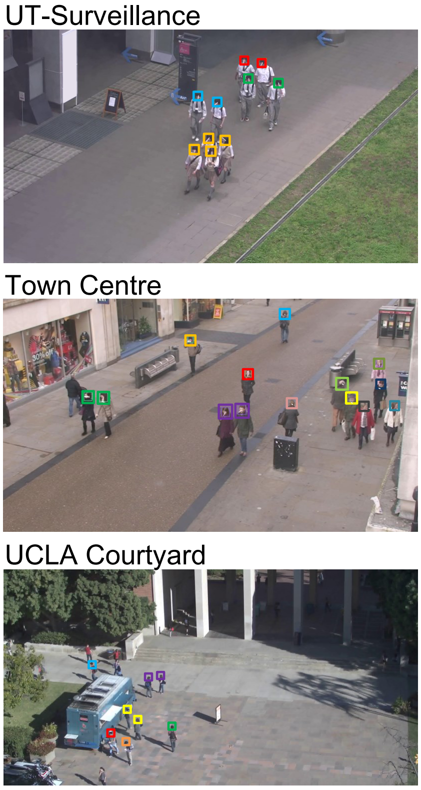
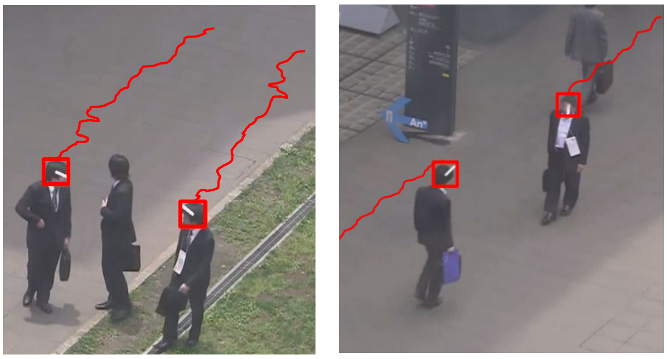


Figure 3.6: Sample frames from sequences used in our work.

ities, such as those waiting in queues to buy food, talking to one another, or walking in groups.

Pedestrian trajectories and head poses were collected from the UT-surveillance dataset using the method described in Subsection 3.2.6. Trajectories provided along with the dataset were used for the Town Centre and UCLA Courtyard dataset, and the head poses were obtained in the same way as the UT-Surveillance dataset. Correctly tracked trajectories were manually annotated with social group IDs. The details on 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 3.1, and example frames in the sequences are in Figure 3.6.

We divided our annotated social groups into three disjoint sets and car-



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

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

Figure 3.7: (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 is stated otherwise in (b).

ried out three-fold cross-validation on the accuracy of estimates to evaluate the performance of our proposed method. Measurements of attention-based cues were calculated and aggregated into cumulative histograms with seven equally-spaced bins ($B_a = 7$), and measurements of position-based cues 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 forest was implemented using the OpenCV library [66] 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 the total training samples. Unless stated otherwise, the clustering threshold T_w was set to 0.7.

3.3.1 Estimation of Group Probability Scores

Group probability scores represent one of the most crucial components of our approach. Accurate estimates of group probability scores results in social groups being correctly clustered. This section explains our evaluation

Table 3.2: Accuracy of estimates from our dataset. Accuracy was measured as average accuracy between two classes to avoid bias problems in test samples. **UTS** stands for UT-Surveillance dataset [71], **TC** stands for Town Centre dataset [38], and **UCLA** for UCLA Courtyard dataset [69]. **Bazzani** indicates results from the approach proposed by Bazzani *et al.* [9]. **P+SVM** presented results using measurements of position-based cues and SVM as classifier, similar to Yamaguchi *et al.* [7]. **P+RT** indicates results using measurements of position-based cues with random forest classifier. **P+A+RT (Proposed)** indicates results with our proposed approach that used measurements of both attention-based and position-based cues with random forest classifier. Note that frame rate in datasets is 30 fps.

Dataset	Approach	N_{past}					
		0	30	60	120	240	∞
UTS	Bazzani	52.5					
	P+SVM	75.1	75.3	75.2	75.3	75.8	75.8
	P+SVM	74.9	75.8	76.5	75.8	76.0	76.4
	P+RT	70.1	73.3	75.9	76.6	78.5	78.1
	P+A+RT (Proposed)	**76.4**	**76.6**	**77.9**	**77.9**	**80.3**	**81.2**
TC	Bazzani	51.4					
	P+SVM	67.3	72.5	73.8	76.3	77.5	78.5
	P+RT	67.6	73.3	75.8	76.2	75.1	76.4
	P+A+RT (Proposed)	**68.3**	**73.9**	**75.4**	**75.2**	**81.4**	**81.8**
UCLA	Bazzani	53.1					
	P+SVM	78.3	79.1	79.1	80.2	81.5	81.2
	P+RT	77.5	78.3	79.0	79.3	81.1	81.4
	P+A+RT (Proposed)	**82.2**	**82.6**	**84.1**	**84.0**	**84.5**	**85.9**

of our approach to calculate group probability scores from both attention-based and position-based cues and compares the results with an approach from previous work. Pedestrian tracks cannot always be accurately obtained from low resolution videos, and the existence of some pairs of pedestrians only overlap for a short time. We therefore conducted tests with different trajectory durations in order to measure how our proposed approach performs in such events.

We compared our method to that by Yamaguchi *et al.* [7] and Bazzani *et*

Table 3.3: Accuracy of estimates using our dataset video downsampled to 0.625 fps.

Dataset	Approach	N_{past}					
		0	1	2	4	8	∞
UTS	P+SVM	75.1	75.3	75.2	75.3	75.8	75.8
	P+A+RT (Proposed)	**75.3**	**77.1**	**78.0**	**77.7**	**78.1**	**79.1**
TC	P+SVM	**67.0**	**67.6**	**69.1**	67.3	67.3	67.3
	P+A+RT (Proposed)	66.2	66.7	68.2	**72.7**	**71.2**	**72.7**
UCLA	P+SVM	77.3	78.0	79.4	80.4	81.1	80.9
	P+A+RT (Proposed)	**80.7**	**81.1**	**81.2**	**81.9**	**82.3**	**82.9**

al. [9] who proposed solving the problem of pairwise social group discovery in a similar setting to that in our approach to estimating group probability scores. Comparisons were done by varying the number of 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. [1] Pedestrians in this test were estimated to be in the same social group if the output from the random forest was more than 0.5 and vice versa. We also conducted experiments on a random forest trained with feature vectors obtained from measurements of position-based cues alone to demonstrate the accuracy of the random forest on position-based cues.

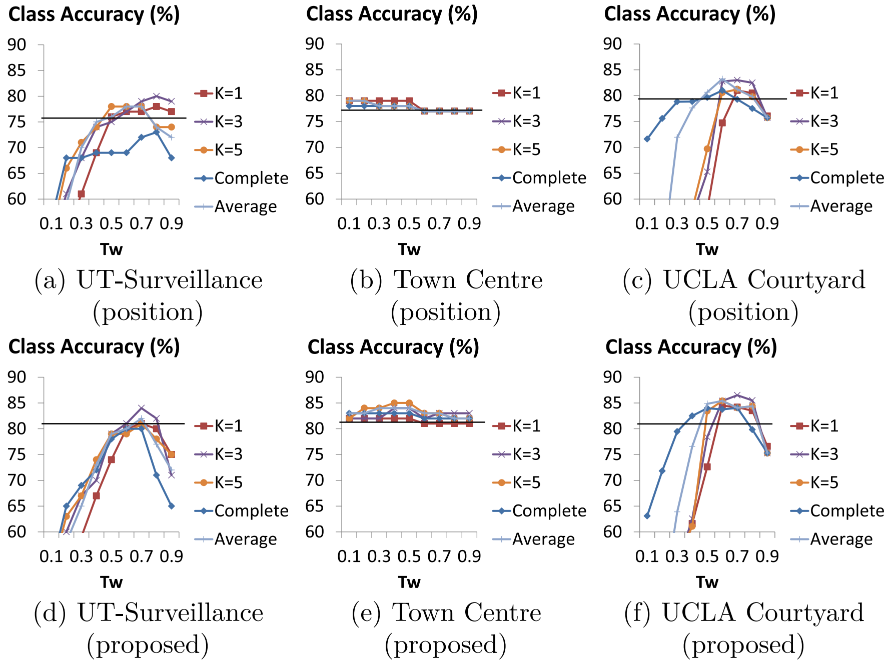
Since the numbers of positive and negative samples were unbalanced in all of the datasets, class accuracy, *i.e.*, the average between the accuracy of each class was used to evaluate accuracy. The results are listed in Table 3.2. It can be seen that the approach by Bazzani *et al.* [9] that assumes stationary pedestrians cannot be applied with our case, and our approach improved

[1] Accuracy of the approach by Bazzani *et al.* [9] were calculated by comparing estimation result of pedestrians in each frame to the ground truth label, and therefore listed in $N_{past} = 0$.

the accuracy of social group discovery tasks in every case. However, it can also be seen that with no past frame information ($N_{past} = 0$), our approach only slightly improved accuracy, but with more available past frames, the improvements from the approach by Yamaguchi *et al.* [7] became 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 social group did not always talk to one another, and therefore the improvements to accuracy were small in cases with low N_{past}. However, attention-based cues can strongly suggest social group relationships even if such cues are rarely observed, *e.g.*, a talking event is a strong indicator of a social group, even if it occurs in a few frames. This makes improvements to accuracy more significant with large N_{past}. However, long tracking trajectories are usually obtained from the tracker in real applications, and situation with low N_{past} are not typical. Therefore, high levels of accuracy can be expected in real situations with the proposed approach. It can also be seen that the accuracy of the random forest estimators trained with position-based cues is comparable to that achieved by Yamaguchi *et al.* [7]. This indicates that the random forest 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.* [7], who used low frame-rate videos. We tested our method with the datasets down-sampled to 0.625 fps and the numbers of available past frames were $N_{past} = 0$, 1, 2, 4, 8, and $N_{past} = \infty$. The results are summarized in Table 3.3. Our approach also did not improve accuracy in low resolution videos with limited numbers of available past frames N_{past}, but it improved accuracy with more available past frames, similarly to that in the previous discussion. This demonstrates that our approach can also be applied to low frame-rate videos, and can greatly improve the accuracy of estimates given that some past frame information is available.

Figure 3.7a has an example of a case where our approach correctly inferred



(a) UT-Surveillance (position)

(b) Town Centre (position)

(c) UCLA Courtyard (position)

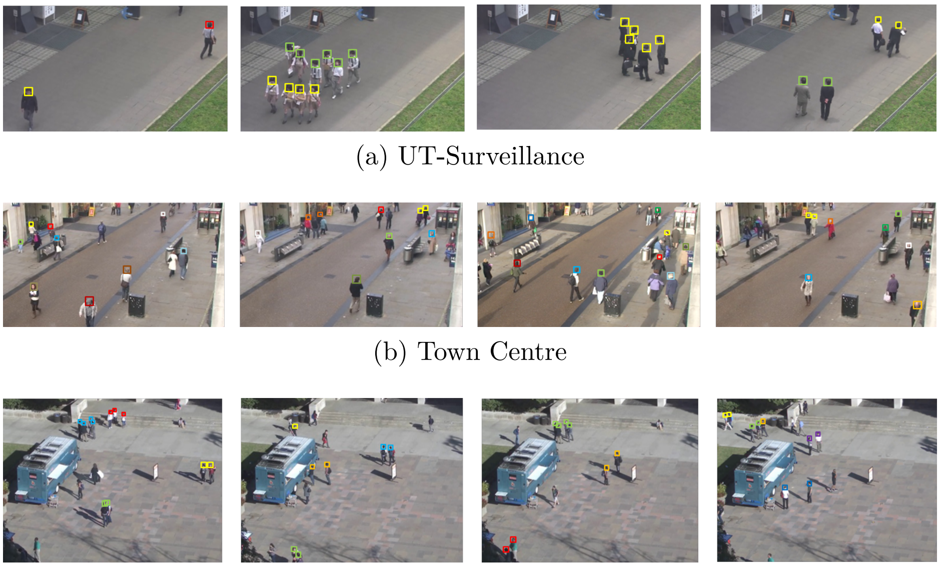
(d) UT-Surveillance (proposed)

(e) Town Centre (proposed)

(f) UCLA Courtyard (proposed)

Figure 3.8: Results obtained from our clustering approach. **Position** indicates that group probability scores used in clustering approach were calculated using measurements of position-based cues and **Proposed** indicates that they are calculated using both attention-based and position-based cues. $\boldsymbol{K = 1}$, $\boldsymbol{K = 3}$, and $\boldsymbol{K = 5}$, indicate top-k linkage criteria with $T_K = 1, 3, 5$, and **Complete** and **Average** indicate complete-linkage and average-linkage criteria, respectively. Straight lines on graph plot the accuracy obtained from pairwise group probability scores without graph clustering. Figure is best viewed in color.

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 did not hold. Figure 3.7b shows an example of such limitations, where the two pedestrians in the same social group are walking toward each other. We assumed that the *gaze exchange* cues were not informative when pedestrians turned their heads in the direction they were walking in. Therefore, although



(a) UT-Surveillance

(b) Town Centre

(c) UCLA Courtyard

Figure 3.9: Examples of results for our clustering approach on three datasets. Images have been cropped for better visibility.

the pedestrians are looking at each other, such information is disregarded and caused our approach to fail in this case. This suggests that more complex assumptions are needed to handle such cases.

3.3.2 Social Group Clustering Test

This section explains our evaluation of the proposed social group clustering algorithm. Accuracy was calculated similarly to the pairwise accuracy in Subsection 3.3.1, except that pedestrians were estimated to be in the same social group if they were clustered into the same social group by the algorithm.

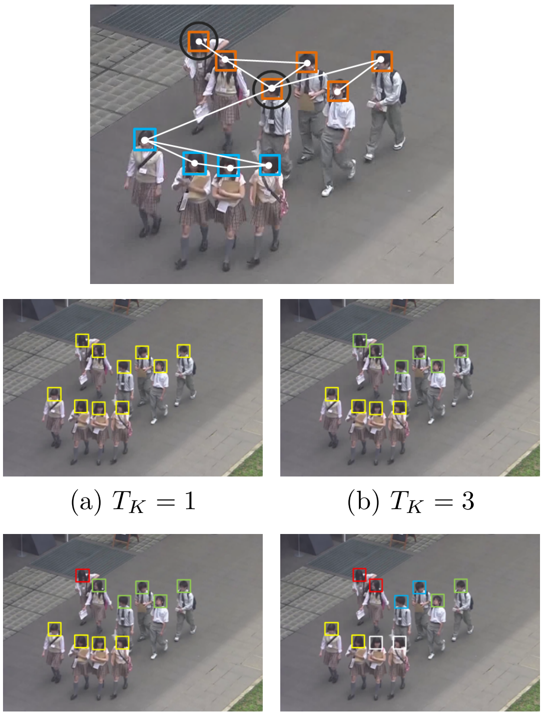
We carried out tests using hierarchical clustering with top-k linkage criteria, and compared the results with those using complete-linkage and average-linkage criteria, which are common criteria used in hierarchical clustering algorithms. Hierarchical clustering with complete-linkage criteria calculates the score for each pair of clusters as the minimum of the weights between each pair of nodes in the clusters, and the one with average-linkage criteria calculates the score as an average of the weights between each pair of nodes in the clusters.

We conducted tests on three datasets with different clustering criteria and clustering threshold T_w to evaluate the clustering approaches. The results are plotted in Figure 3.8. **Position** indicates the accuracy of the graph clustering approach with group probability scores calculated using measurements of position-based cues. **Proposed** indicates the accuracy of our proposed approach which used measurements of both attention-based and position-based cues to calculate the group probability scores. The $\boldsymbol{K = 1}$, $\boldsymbol{K = 3}$, and $\boldsymbol{K = 5}$ indicate the accuracy of the graph clustering approach with top-k linkage criteria with $T_k = 1$, $T_k = 3$ and $T_k = 5$, respectively. **Complete** and **Average** indicate the accuracy of the graph clustering approach with complete-linkage and average-linkage criteria, respectively. The straight lines on the graph plot the accuracy obtained from pairwise group probability scores without graph clustering. It can be seen that in the UT-Surveillance dataset and UCLA Courtyard dataset, the approach yields highest accuracy with $T_K = 3$, while there are no significant changes in accuracy by varying T_K in the Town Centre dataset. This is because while the top-k linkage criteria yielded little improvements in scenes with small social groups such as in the Town Centre dataset, it was crucial for distinguishing large groups of pedestrians in the UT-surveillance dataset. The results also revealed that the highest accuracy was achieved with top-k linkage criteria with $T_w = 0.7$ and $T_K = 3$ in both UT-Surveillance and UCLA Courtyard datasets. Examples of results for our clustering approach on three datasets are in Figure 3.9.

Clustering results from the same scene with different approaches are given in Figure 3.10. Squares with same color in top image indicate ground truth social groups. Edges with group probability scores $w_{ij} > 0.7$ are indicated by white lines in figure, and black circles indicate a pair of pedestrians with group probability score of 0.3. Subfigures (a)-(d) show results obtained from our approach using top-k linkage criteria with (a) $T_K = 1$, (b) $T_K = 3$, (c) average-linkage criteria and (d) complete-linkage criteria.

It can be seen from the figure that although the pair of pedestrians in the black circles has a low group probability score, it is correctly estimated to be in the same social group by top-k linkage criteria. This demonstrates the effectiveness of top-k linkage criteria in discovering social groups. It can also be seen from the figure that with $T_K = 1$, two different social groups are mis-detected as being the same social group by an outlier link between them. With $T_K = 3$, the top three group probability scores are used to calculate the score and make the approach robust to the outlier link. With average-linkage criteria, the leftmost pedestrian in the top social group is mis-detected as being in a different social group due to the low probability score with the rest of the pedestrians in the social group. Further, complete-linkage criteria separated pedestrians from the same social group due to the low probability scores of some pairs of pedestrians. It can be seen that with appropriate T_K, the best clustering results can be achieved with top-k linkage criteria.

Figure 3.11 presents the results obtained from our approach with different linkage criteria and clustering threshold T_w. Pedestrians in the same social group are separated with $T_w = 0.9$ because the confidence threshold is too high causing the algorithm to prematurely terminate. With $T_w = 0.7$, single-linkage criteria ($T_K = 1$) grouped two pedestrians in the middle with the social group at the bottom of the scene. The social groups are correctly detected with top-k linkage criteria with $T_K = 3$. Average-linkage criteria correctly detected the social groups with $T_w = 0.5$, but failed to detect a pedestrian with $T_w = 0.7$. This also demonstrates that we can achieve correct clustering results with appropriate T_K and T_w. However, it can also be



(a) $T_K = 1$

(b) $T_K = 3$

(c) Average

(d) Complete

Figure 3.10: Example results obtained from our clustering approach. Squares with same color in top image indicate ground truth social groups. Edges with group probability scores $w_{ij} > 0.7$ are indicated by white lines in figure, and black circles indicate a pair of pedestrians with group probability score of 0.3. Subfigures (a)-(d) show results obtained from our approach using top-k linkage criteria with (a) $T_K = 1$, (b) $T_K = 3$, (c) average-linkage criteria and (d) complete-linkage criteria. Figure is best viewed in color.

seen that average-linkage criteria also yielded correct results with $T_w = 0.5$ and this suggests that the criteria might be more effective under certain circumstances. This also suggests further investigations are needed to determine the properties of social groups and to find which linkage criteria to apply to such social groups.

3.3.3 Analysis of Attention-based Cues

We carried out two tests to measure how each attention-based cue contributed to the accuracy of the proposed approach and the relative importance of these cues when applied to different scenes.

The first test analyzed the variable importance matrix obtained from the random forest. The variable importance measured the impact of each feature in the training feature vectors. The importance of the $j-th$ feature was calculated by permuting the value of the $j-th$ feature in the training data. The difference in out-of-bag (OOB) error before and after permutation was then compared. Features that produced large differences in errors were ranked as being more important than features that produced small differences.

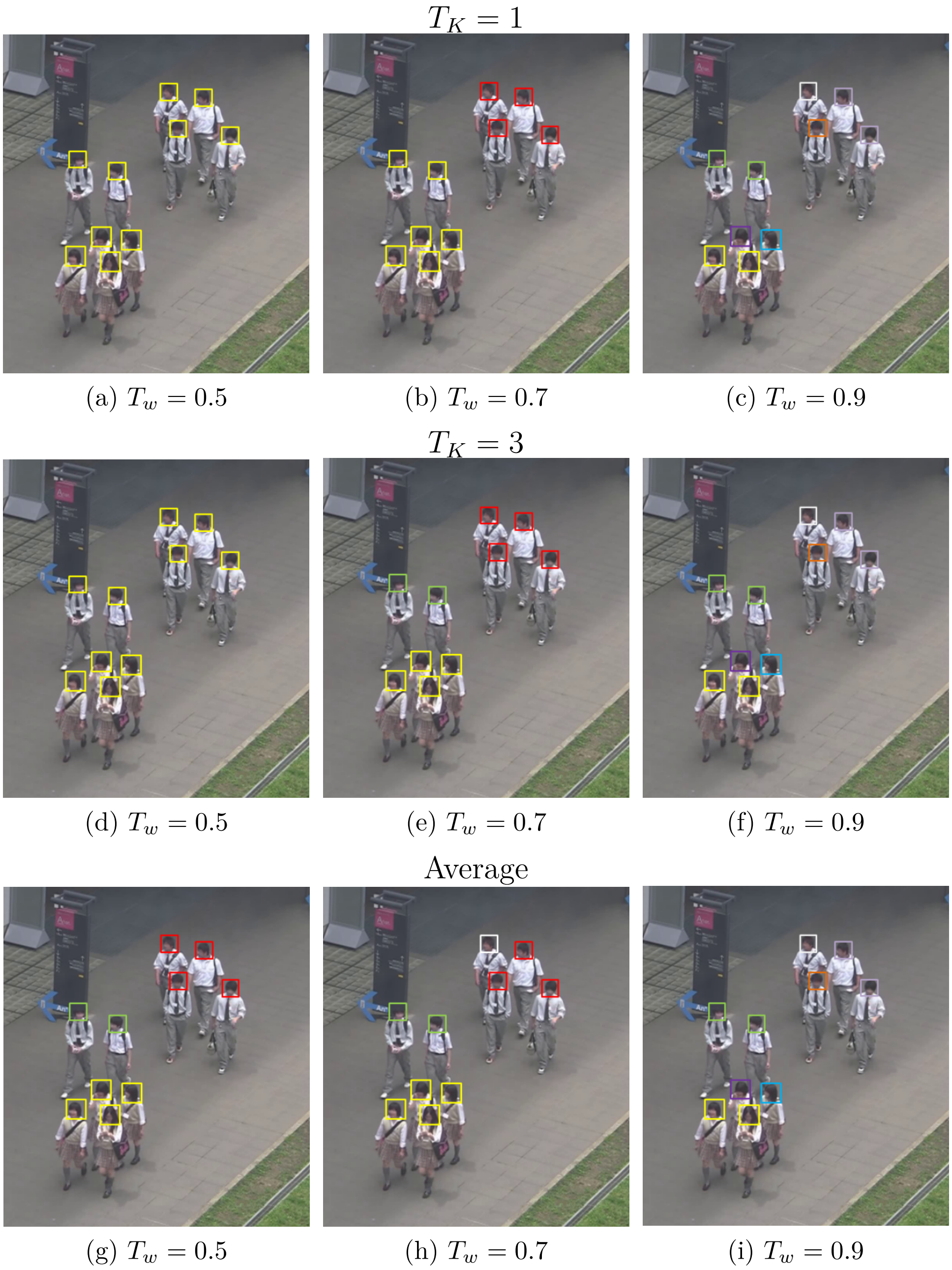
The results are plotted in Figure 3.12. The normalized average values of the importance of all bins of each feature are shown. It can be seen that in the UT-Surveillance and UCLA Courtyard dataset, the *gaze exchange* cue has the highest importance values amongst the attention-based cues. This is intuitive because several large groups were observed in the UT-Surveillance scene. As conversation events usually occurred in large groups, it made the *gaze exchange* cue powerful in determining social groups. Pedestrians standing still who were having conversations were observed in the UCLA Courtyard dataset. Position-based cues could not distinguish pedestrians from different groups standing close together, and therefore the *gaze exchange* cue greatly helped in detecting these groups. Furthermore, pedestrians in this video walked with relatively similar speeds and directions, making attention-based cues crucial to discovering social groups in this scene. The importance of the *mutual attention* cue was higher than other attention-based cues in the Town Centre dataset, where only small groups were observed. This can be interpreted as the *mutual attention* cue, which is measured based on the assumption that pedestrians from different groups usually looked in different directions, contributed more to separating pedestrians who were not from the same social group than the *gaze exchange* cue, which helped to detect pedestrians from the same social group as was done in the UT-Surveillance dataset.

We evaluated the importance of each feature in the second test by calculating the accuracy of our proposed approach by omitting measurements from the feature vector. The results are plotted in the graphs in Figure 3.13. Similar results to those from the previous tests can be observed. The greatest drop in accuracy is observed when *mutual attention* and *gaze exchange* measurements have been omitted and this suggests that these cues are important in the task of social group discovery in all scenes.

It can be seen from the results that attention-based cues greatly helped in discovering social groups in our task. The importance of attention-based cues varied with different kinds of scenes. The *gaze exchange* cue contributed more to scenes with pedestrians who were standing still, where position-based cues would have failed to discover groups due to limited information on movements of pedestrians. The cue was also important in scenes with large social groups, which usually had more conversation events than smaller ones. The *mutual attention* cue helped to separate pedestrians who were not in the same social group, especially in scenes with many small social groups. The *walking focus* and *walking speed* cues generally helped to increase the accuracy in every case. This suggests that future work is to develop an approach that can assign weights to cues in order to focus on discovering specific kinds of social groups. For example, assigning more weight to the *gaze exchange* cue could help tasks that were aimed at discovering large social groups, while more weight on *mutual attention* could help tasks that were aimed at discovering pedestrians who did not belong to any groups.

3.4 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 social 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 only used measurements of position-based cues. We believe that there are still other cues humans can use to discover social groups, and investigating and discovering these cues will be important in future work.



$T_K = 1$

(a) $T_w = 0.5$ (b) $T_w = 0.7$ (c) $T_w = 0.9$

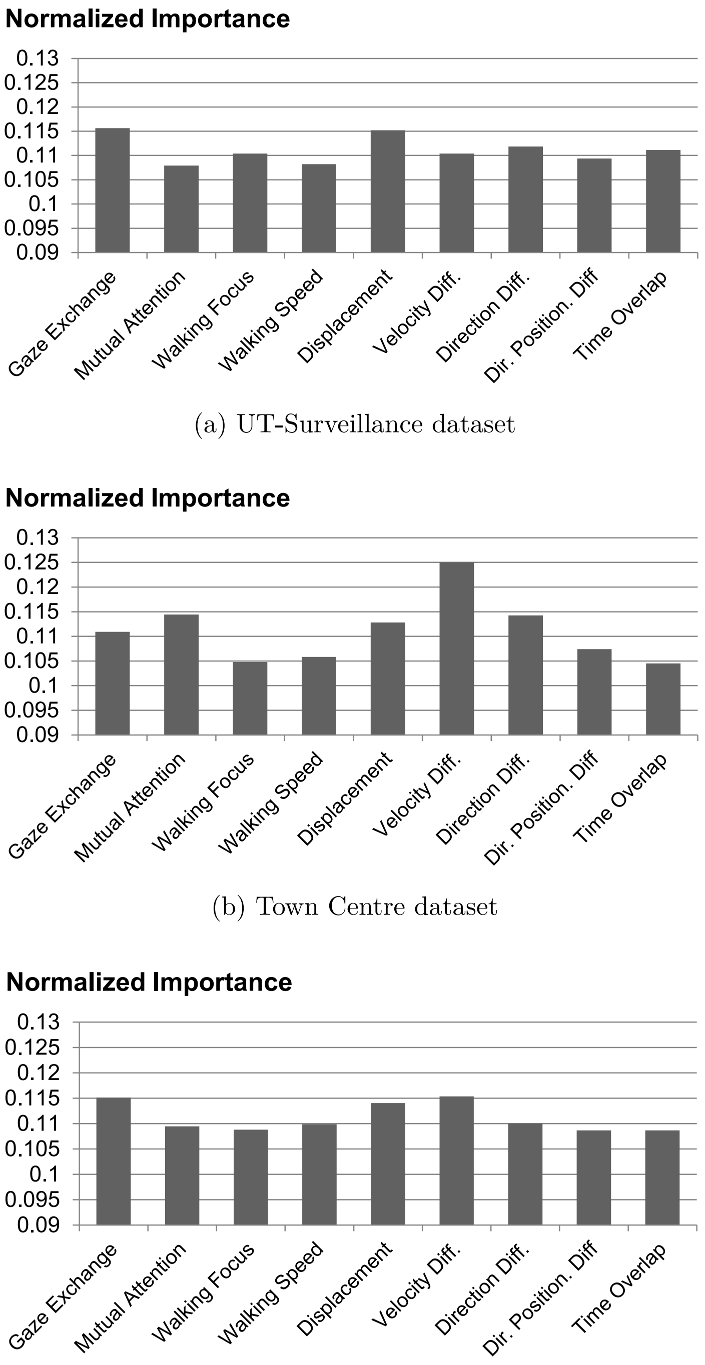
$T_K = 3$

(d) $T_w = 0.5$ (e) $T_w = 0.7$ (f) $T_w = 0.9$

Average

(g) $T_w = 0.5$ (h) $T_w = 0.7$ (i) $T_w = 0.9$

Figure 3.11: Results from our approach with different linkage criteria and clustering threshold T_w. Three rows correspond to result with top-1, top-3, and average-linkage criteria, respectively. Three columns correspond to results with $T_w = 0.5$, $T_w = 0.7$, and $T_w = 0.9$, respectively. Figure is best viewed in color.

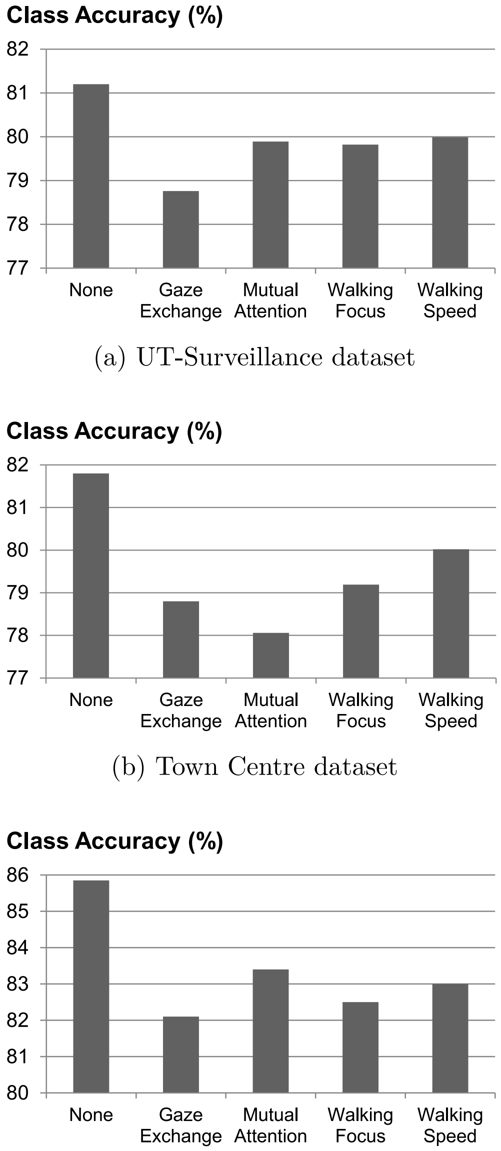


(a) UT-Surveillance dataset

(b) Town Centre dataset

(c) UCLA Courtyard dataset

Figure 3.12: Normalized variable importance of measurements. Average value of variable importance of each bin is shown.



(a) UT-Surveillance dataset

(b) Town Centre dataset

(c) UCLA Courtyard dataset

Figure 3.13: Accuracy calculated by omitting measurements of attention-based cues from proposed approach. Names below columns indicate omitted measurements.

Chapter 4

Unsupervised Social Group Type Discovery

4.1 Introduction

One aspect of social group analysis is how humans interact within the group. Groups of pedestrians standing near a shop stand can imply their interests in products at the shop stand, while those who talk to each other while walking might imply their lack of interests in those products. However, manual observation of such human behaviors in lengthy videos is time consuming and almost impossible to be conducted. Moreover, important social groups types are not always obvious. Pedestrian groups knowing each other slow down when they walk past, or those who are choosing what to buy in the convenience store walk slowly near the stands they are interested in, and labeling these groups is not a trivial task.

In the previous chapter, it is shown that various social group behaviors can be modeled with attention-based and position-based cues. Therefore, we expect to be able to find important insights on what types of social groups are found in each scene by examining outputs from unsupervised social group type discovery which take into account these cues.

In real life, social groups are not always rigid; one social group in a frame can be separated into multiple groups involving in different actions in the subsequent frames, or two different groups can be merged and conduct the same action. However, social groups discovered by many approaches [41, 8, 12, 7, 15] including our previous work in Chapter 3 assumed social groups to be rigid and do not take into account this fact, therefore, we proposed to include transient group constraint to our approach. Groups with transient constraints can change over time, either one group can split into two groups, and two groups can merge into one group over time, and this constraint allows for more realistic applications. Bazzani *et al.* [13] proposed an approach that can handle merging and splitting of groups by using decentralized particle filtering. However, the groups are modeled as pedestrians with similar walking speed and direction, and did not take into account the attention of each pedestrians.

We proposed an approach to discover transient social groups in video in real time based on our previous work in Chapter 3. The simplest way to cope with transient constraint is to discover social groups at each time unit separately. Discovering group in this manner, however, is prone to erroneous results due to noisy input data, causing groups to frequently shift. Therefore, we applied evolutionary clustering approach proposed by [72] in order to consider group structure from the previous time unit in order to filter out noisy results.

4.2 Proposed Framework

In order to discover types of social groups, pedestrians are first clustered into groups in each frame using an evolutionary clustering approach proposed by [72]. This approach is based on hierarchical clustering with an additional

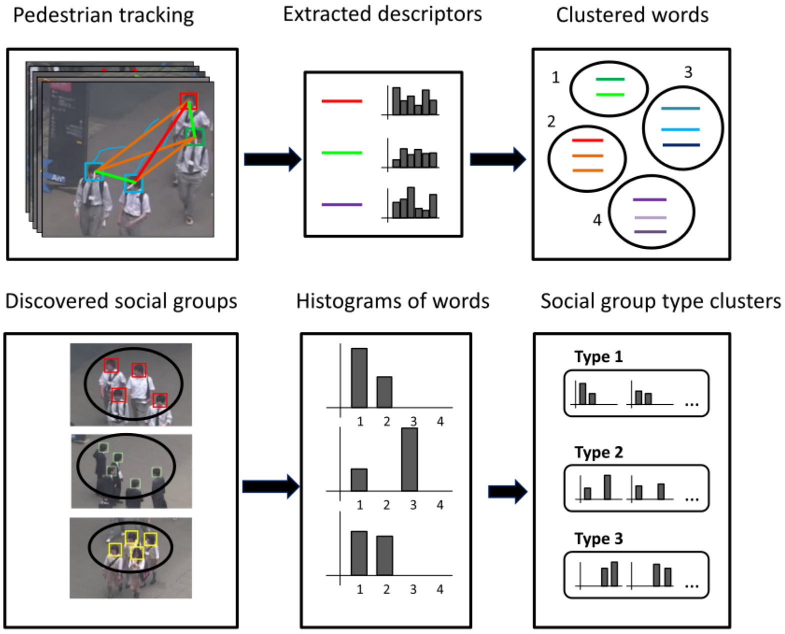


Figure 4.1: Proposed framework. Pedestrians are first clustered into groups in each frame using an approach similar to that in the previous chapter. Descriptors for each pair of pedestrians in the group are collected over the entire video and clustered into distinct visual words. A normalized histogram measuring the frequency of visual words in each group are then calculated and the histograms are collected from social groups over the entire video are then clustered into multiple social group types.

approach to preserve group structures from previous time step. With the groups discovered, we apply bag-of-features-based approach to cluster the social groups into multiple types. Descriptors for each pair of pedestrians in the group are collected over the entire video and clustered into distinct visual words. A normalized histogram measuring the frequency of visual words in each group are then calculated and the histograms are collected from social groups over the entire video are then clustered into multiple social group types. The framework for our method is outlined in Figure 4.1.

4.2.1 Social Group Discovery with Temporal Smoothness

To take into account decomposible social groups, we estimate social group memberships separately for each time unit t. At each time unit t, we collect measurement descriptors similar to those used in Chapter 3 within the γ-sized time window centered at t, *i.e.*, collect measurements from time unit $t_w \in [t - \frac{\gamma}{2}, t + \frac{\gamma}{2}]$. Histograms of measurements for each pedestrian are then calculated and concatenated into a feature vector, and the feature vectors collected from train data is used to train a random forest regressor. In test scene, this regressor is applied to the feature vector between each pair of pedestrians to calculate the group probability score, estimating the likelihood that two pedestrians belong to the same group. We refer readers to Chapter 3 for more details.

For each time unit t, we obtain a graph $K = (V, E)$, where graph nodes $V = \{v_1, v_2, \ldots, v_n\}$ where n is the number of pedestrians in the scene, represents pedestrians in the scene. Edges E represents relationships between pedestrians. Each edge e connecting two nodes v_i and v_j is associated with group probability score w_{ij} specifying the likelihood that their corresponding pedestrians are from the same group. We perform hierarchical clustering in order to assign groups G_i to each node v_i. The clustering steps are similar to the one used in Chapter 3. The algorithm starts with each node v_i in their own cluster. The algorithm then merges two nodes with maximum merge benefit in each iteration until the maximum merge benefit falls below a threshold T_w.

In the standard hierarchical clustering approaches, the merge benefit between two clusters only considers the weights of their members. However, we applied the evolutionary clustering [72] in order to take into account group structure from the previous time unit. Evolutionary clustering is an approach for clustering while considering the previous cluster structures from the previous frames, and the paper proposed several heuristics for calculating merge benefits in order to merge two clusters. In our work, we applied the heuristic stated to be the best from the paper to calculate merge benefit $mb(m)$, which is calculated as

$$\begin{aligned} mb(m) = sim(m) - cp \cdot \underset{\substack{v_i \in leaf(m_l) \\ v_j \in leaf(m_r)}}{E} (d_{K'}(v_i, v_j)) - (d_K^m(v_i, v_j)) \qquad (4.1) \\ +cp \cdot \underset{\substack{v_i \in leaf(m) \\ v_j \notin leaf(m)}}{E} (d_{K'}(v_i, v_j)) - (d_K^m(v_i, v_j)), \qquad (4.2) \end{aligned}$$

where m is the merge point of the two clusters, where m_l and m_r are the two clusters being considered. $sim(m)$ is the similarity of two clusters, calculated using k-top linkage criteria. cp is a threshold specifying how past clusters affect the result of current cluster. $d_K(v_i, v_j)$ and $d_{K'}(v_i, v_j)$ are the tree distance between node v_i and v_j, respectively.

The first term, $sim(m)$, measures the similarity of two clusters as in standard hierarchical clustering. Clusters of pedestrians with high group probability score will have high $sim(m)$. The second term, $\underset{\substack{v_i \in leaf(m_l) \\ v_j \in leaf(m_r)}}{E} (d_{K'}(v_i, v_j)) - (d_K^m(v_i, v_j))$, calculates the similarity between current clustering and the clustering result in the previous frame. Tree distance between each node v_i in the first cluster m_l, and each node j the second cluster, m_r, is calculated as $d_K^m(v_i, v_j)$. The tree distance $d_K^m(v_i, v_j)$ is calculated by counting the number of steps required to traverse from node v_i to v_j if m_l and m_r is merged. This distance is compared to the previous clustering result, $d_{K'}(v_i, v_j)$. If the distances are similar, merging at m is more likely to be correct. The last term, $\underset{\substack{v_i \in leaf(m) \\ v_j \notin leaf(m)}}{E} (d_{K'}(v_i, v_j)) - (d_K^m(v_i, v_j))$, calculates the benefit if the other nodes are merged instead. Tree distance between merged nodes $v_i \in leaf(m)$ and unmerged nodes $v_j \notin leaf(m)$ is calculated and compared to the distance

Algorithm 2 Social group clustering with single-linkage criteria

```
Input: Graph $K = (V, E)$, group probability scores $W$, a clustering threshold $T_w$, the clustering result from previous frame $K' = (V', E')$ and a threshold $cp$.
Output: Group assignments $\boldsymbol{G} = \{G_1, G_2, \dots, G_N\}$
  $\boldsymbol{G} := \{G_1, G_2, \dots, G_{|V|}\}$, where $G_i = \{v_i\}$
  while $|\boldsymbol{G}| > 1$ do
    $G_i, G_j := \operatorname{argmax}_{G_i, G_j \in \boldsymbol{G}} mb(G_i, G_j, K')$)
    $w_{max} := \max(mb(G_i, G_j, K'))$
    if $w_{max} < T_w$ then
      return $\boldsymbol{G}$
    else
      Remove $G_i$ and $G_j$ from $\boldsymbol{G}$
      Add $G_i \cup G_j$ to $\boldsymbol{G}$
    end if
  end while
```

between these two nodes in the past. Similar distance from v_i to v_j implies that the cluster containing v_i and the cluster containing v_j should be merged in this step, and impose the penalty of not doing so. The pseudo-code for this algorithm is given in Algorithm 2.

The clustering threshold, T_w, which determines when to stop the clustering approach, plays an important role in determining the accuracy of our approach. However, it is not trivial to set the threshold, and therefore we propose to select an appropriate T_w from the train data. In order to select an appropriate T_w, we perform grid searching approach which select values $\hat{T_w}$ from a search grid with a certain interval, *e.g.*, $\hat{T_w} \in \{0.0, 0.1, \dots, 1.0\}$ and conducted N-fold cross validation with train data using $\hat{T_w}$ as clustering threshold. The value of $\hat{T_w}$ that yields maximum cross validation accuracy is expected to provide maximum accuracy with test data, and thus selected as T_w.

With this algorithm, we obtain an estimation of pedestrian groups $\boldsymbol{G} = \{G_1, G_2, \dots, G_N\}$, where $G_i = \{q_{i1}, q_{i2}, \dots, q_{im}\}$ denotes a group of pedestrians and $\boldsymbol{Q} = \{q_1, q_2, \ldots, q_n\}$ denotes a set of pedestrians in the scene at each time unit t.

4.2.2 Unsupervised Social Group Type Discovery

In this section, we describe our approach to cluster pedestrian group types by the interactions of their members. A group of pedestrians standing and talking to each other and pedestrians walking together can be separated by their movement speed and their attention. However, it is not clear what cues and what are the thresholds needed to separate these groups. Therefore, we use an approach similar to bag-of-features model [73]. This approach first apply an unsupervised clustering is applied to cluster the interactions between pedestrians into multiple types. Each group in the video is then described by the distribution of these interaction types, and the distribution of groups in the video are collected and then further clustered into multiple group types.

We first collect a feature vector from each pair of pedestrians $(q_i, q_j) \in Q \times Q$ from each time unit. The feature vectors from the entire video were collected and clustered using K-means algorithm into T_{nc} clusters $\boldsymbol{C} = \{C_i\}, i = 1, \cdots, T_{nc}$, where T_{nc} is a constant. These clusters represent distinct types of interaction between pedestrians, and are treated as a feature in the bag-of-features model.

At each time unit t, the cluster c_{ij} of the feature vector captured from pair of pedestrians (q_i, q_j) is determined. For each group G_k discovered at time t, the distribution of clusters c_{ij} of pedestrians in the group $(q_i, q_j) \in G_k$ are calculated as a histogram h_k. The histograms $\boldsymbol{H} = \{h\}$ is then collected throughout the video and finally clustered into T_{gt} group types using K-means clustering, where T_{gt} is a constant for the number of group type clusters. In this manner, groups are clustered into distinct types in an automatic manner.

4.3 Experimental Results

We conducted experiments by using a subset of the UT-Surveillance sequence with the length of 125 minutes. The groups are re-labeled to incorporate transient groups that can change over time.

Pedestrian trajectories were collected from the UT-surveillance dataset using the method described in Subsection 3.2.6, and the head poses were obtained in the same way as the UT-Surveillance dataset. Correctly tracked trajectories were manually annotated with social group IDs. We collected a total number of 251 pedestrians and labeled a total of 87 groups.

4.3.1 Social Group Clustering with Temporal Smoothness

To evaluate the social group clustering approach, 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. The parameters were set similarly to the previous chapter ($B_a = 7$, $B_p = 7$), and the random forest was set as follows: 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 the total training samples.

In each frame in test scene, social group clustering was applied and group membership of each pedestrian pairs was then tested and compared with the ground truth and the result is collected throughout the video. Similar to the previous chapter, class accuracy, *i.e.*, the average between the accuracy of each class, was used to evaluate accuracy. The results are shown in Figure 4.2. The figure shows results of our approaches with different cp values. Small stars on the graph show the value of T_w obtained by performing three-fold cross-validation on the train dataset. We also performed tests with previous approach using this dataset shown as *Static*. Due to its inability to

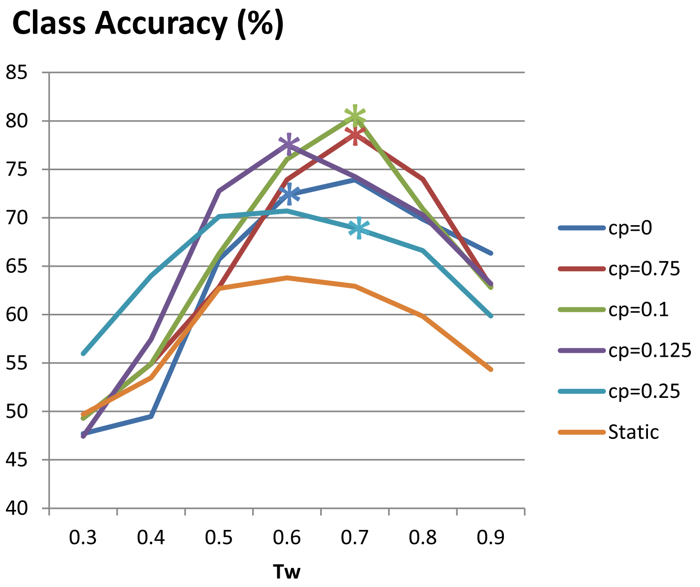


Figure 4.2: Group discovery results of our approaches with different *cp* values. *Static* shows the result using the approach from Chapter 3. Small stars on the graph show the value of T_w obtained by performing three-fold cross-validation on the train dataset. Class accuracy was used to evaluate accuracy.

handle multiple social groups for one pedestrian, the social group that each pedestrian spent the most time in is selected as his social group for training the estimator. It can be seen that the previous approach underperformed when applying to transient group environment, and the approach yields most accuracy with values $cp = 0.1$ and $T_w = 0.7$. This shows that using information from previous frames not only increase the smoothness of the results, but also increases the estimation accuracy. The results also show that our approach correctly selects an appropriate T_w for most of the cp values.

An example of the results is shown in Figure 4.3. The figure shows clustering results of consecutive frames in the video using our approach with $cp = 0$ and $cp = 0.1$. I can be seen that using the history cluster structures help to make the result clusters smooth and less prone to error.

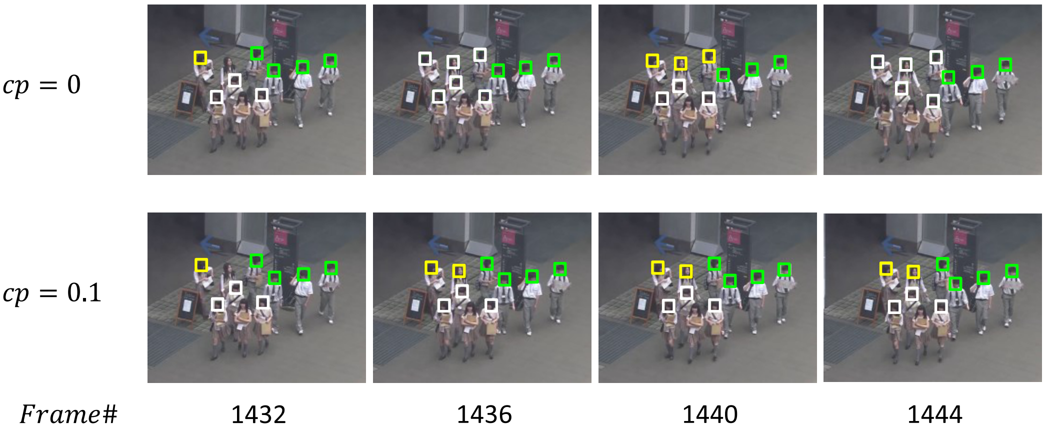


Figure 4.3: Clustering results of consecutive frames in the video using our approach with $cp = 0$ and $cp = 0.125$, respectively. I can be seen that using the history cluster structures help to make the result clusters smooth and less prone to error.

4.3.2 Social Group Type Clustering Test

We performed two tests to evaluate the social group type clustering. The first test compared the results to that of human clustering. The annotator was asked to divide social groups found in the dataset into a number of types which served as a ground truth for the comparison with discovered types. The second test was visual evaluation by human to interpret the results.

The first test was done by asking the annotator to divide pedestrian groups into 3 types, groups of pedestrians walking without talking, groups of pedestrians talking while walking, and groups of pedestrians standing and talking to each other. We set $T_{nc} = 10$ and $T_{gt} = 3$ and applied the algorithm with the social groups obtained from the discovery approach using $cp = 0.1$ and $T_w = 0.7$. In each frame, social group clustering was applied to cluster the pedestrians. We constructed a confusion matrix between the estimated and the annotated group types in order to evaluate the result. Because clustering was done in an unsupervised manner, we aligned the matrix so that maximum diagonal values were maximized. The confusion matrix is

Table 4.1: Confusion matrix between ground truth and estimated group types. Estimated type labels were given by human and are aligned to maximize values along the diagonal line. **W** indicates groups of pedestrians walking without talking, **S** indicates groups of pedestrians standing and talking to each other, and **WT** indicates groups of pedestrians talking while walking.

		Estimated type		
		W	S	WT
Ground truth	W	0.87	0	0.13
	S	0.12	0.89	0
	WT	0.49	0	0.51

Table 4.2: The average measurement values of each discovered visual word. Values from z_1 to z_{10} describe visual words, and the columns from left to right show the averaged measurement values for distance between pedestrians (p_1), velocity difference (p_2), walking direction difference (p_3), difference between walking direction and relative position (p_4), gaze exchange (a_1), head pose difference (a_2), walking focus (a_3), and velocity (a_4), respectively. Measurements are scaled to $[0, 1]$. Cells with high values are colored green, while cells with low values are colored red. This table is best viewed in color.

	$dist$	Δvel	Δdir	$dirpos$	$gaze$	$\Delta pose$	$focus$	$velocity$
z_1	0.17	0.07	0.19	0.25	0.44	0.24	0.35	0.11
z_2	0.20	0.10	0.04	0.37	0.45	0.19	0.30	0.52
z_3	0.19	0.07	0.05	0.70	0.42	0.20	0.30	0.41
z_4	0.28	0.14	0.08	0.43	0.35	0.57	0.46	0.48
z_5	0.18	0.12	0.04	0.06	0.43	0.26	0.38	0.46
z_6	0.57	0.15	0.05	0.46	0.44	0.21	0.29	0.54
z_7	0.56	0.14	0.05	0.10	0.44	0.22	0.30	0.43
z_8	0.16	0.05	0.29	0.27	0.39	0.67	0.31	0.05
z_9	0.47	0.19	0.22	0.13	0.32	0.69	0.36	0.23
z_{10}	0.83	0.39	0.28	0.37	0.41	0.63	0.39	0.48

displayed in Table 4.1. The average feature values of each discovered visual word are displayed in Table 4.2, and the distribution of visual words in each discovered group type is shown in Table 4.3.

Table 4.3: The distribution of visual words in each discovered social group type. Ten columns correspond to the ratio of each visual word in each discovered group type. Major visual words in each group type are colored green.

	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	z_{10}
W	0.01	0.47	0.51	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.77	0.08	0.00
WT	0.01	0.08	0.04	0.20	0.58	0.00	0.01	0.00	0.05	0.02

It can be seen that groups of pedestrians walking without talking (W) and standing pedestrians (S) are clustered correctly with more than 80% accuracy. However, as talking event did not always occur even in the WT group, they were clustered into W in frames where pedestrians in this group are not talking. This can be thought of a more detailed labeling of the social group and might be preferable depending on the application. This suggests more sophisticated tests might need to be performed in order to analyze these groups.

It can also be seen that the distributions of discovered pedestrian groups follow our assumption of social groups. Walking pedestrians (W) are described mainly by visual words z_2 and z_3, which describe pedestrians with low relative distance, high velocity and low walking direction difference. Standing pedestrian groups (S) are explained by visual word z_8 with low relative distance, low velocity, and high gaze exchange activity (indicated by low gaze exchange value). Pedestrians who talk while walking (WT) are mainly described by visual words z_4 and z_5, which signified pedestrians with low relative distance, high velocity, and high gaze exchange activity.

The second test is performed by setting $T_{nc} = 10$ and $T_{gt} = 4$ and ask human annotator to interpret the meaning of each social group type. The results are shown in Figure 4.4. Three discovered social group types were identified similarly as in the previous section, which are groups of pedestrians walking without talking, groups of pedestrians talking while walking, and

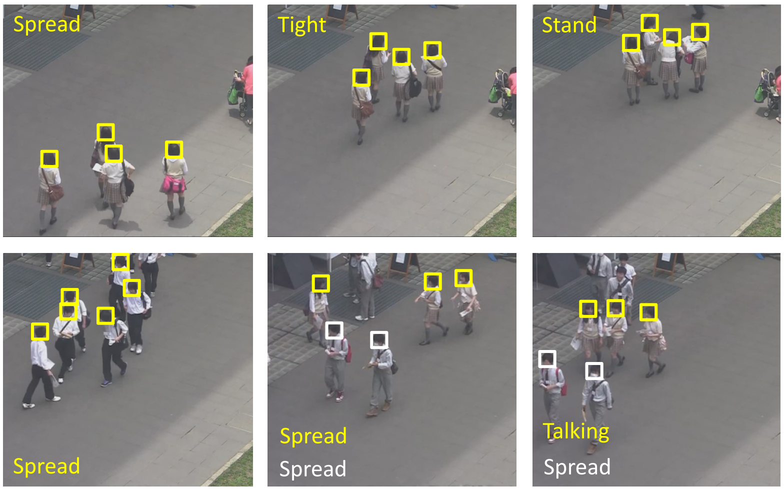


Figure 4.4: Examples of results obtained by the proposed approach with $T_{nc} = 10$ and $T_{gt} = 4$. Social groups are represented by colored blocks, and the text with the same color as the boxes states the estimated group type for the corresponding social group. **Spread** shows pedestrian groups walking in a spread-out manner, **Tight** shows pedestrian groups walking more tightly, **Stand** shows pedestrian groups standing and talking, and **Talking** shows pedestrian groups talking while walking.

groups of pedestrians standing and talking to each other. The fourth social group type is identified by human as groups of people walking in a spread-out manner. In the figure, social groups are represented by colored blocks, and the text with the same color as the boxes states the estimated group type for the corresponding social group. **Spread** shows pedestrian groups walking in a spread-out manner, **Tight** shows pedestrian groups walking more tightly, **Stand** shows pedestrian groups standing and talking, and **Talking** shows pedestrian groups talking while walking. It can be seen that a potentially important social group types can be identified by this approach. Setting T_{gt} to values more than 4, however, yields no interpretable results. However, we believe that applying this to other type of scenes will yield different results and this approach could discover potentially important groups in that scene.

4.4 Conclusion

We proposed an unsupervised social group type discovery approach that can operate without human labels. Our approach has the key idea that human social groups can be categorized by analyzing interaction of pedestrians within the group. The experimental results show that our approach can discover several important group types without any prior labels. We also proposed an approach to discover social groups with transient property. Social groups are discovered separately in each frame with previous social group structures taken into account to help removing outliers. The experimental results confirmed that the accuracy is significantly improved by factoring in the social group structures from previous frame.

We believe that there are still various feature types that can possibly discover important groups that are not covered in our work, and there are still various clustering approaches that can yield better results. There are numerous ways to improve each component of this framework, and the it is our goal to seek the perfect set of components that can achieve the best results.

Chapter 5

Conclusions

5.1 Summary

This thesis demonstrated approaches to discover social group and social group types by using attention-based and position-based cues. Attention-based cues are derived from human attention estimated from their head pose, while position-based cues are acquired from spatial information of the pedestrians acquired from pedestrian tracking. We first described an unsupervised approach to acquire head pose for pedestrians in the scene without manual labeling process. A group discovery approach is then established by training a random forest with attention-based and position-based measurements. Our approach demonstrated that human social groups can be robustly estimated and the use of attention-based cues increases the accuracy over using position-based cues alone. We also utilized attention-based and position-based cues to discover social group types in an unsupervised manner with the key idea that human social groups can be categorized by analyzing interaction of pedestrians within the group. Both quantitative and qualitative experiments show that important social group types can be categorized in an unsupervised manner by using these cues.

This paper shows that social group information as well as different types of social groups in the video can be robustly acquired with our proposed approach. This would surely enrich current social group analysis processes and makes it possible for more detailed and real-time analyses, which would open up new possibilities for vast array of applications.

5.2 Contributions

Main contributions of this thesis are summarized as follows.

Unsupervised head pose estimation

We proposed an unsupervised head pose estimation approach that is based on two key ideas: automatically acquiring a training image dataset with ground-truth head pose and segmenting a scene into multiple regions with similar head appearances. This is the first work to use walking directions as a cue to infer head pose and solve head pose estimation task in an unsupervised manner. The accuracy is further improved by our approach to handle large variations that occur in head appearance within the same scene. The whole method is fully unsupervised, which is a significant advantage when applied to practical scenarios.

Social group discovery using attention-based cues

We proposed an approach to discover social groups by using attention-based and position-based cues. While attention-based cues were applied with several researches in computer vision, we were the first work to apply these cues to social group discovery. Attention-based cues are modeled as a set of features based on human attention, and these features are used to learn a set of decision trees that represent the behaviors of social groups. Experimental results show that social groups can be robustly estimated and the use of attention-based cues increases the accuracy over using position-based cues alone

Unsupervised social group type discovery by modeling human interactions

We propose a group discovery approach which takes into account the transient group property and an evolutionary clustering approach is applied removing erroneous results. The discovered social groups are then categorized by analyzing interactions among their members in an unsupervised manners. Our approach is the first that propose to categorize social groups by examining the interactions between their members. Both quantitative and qualitative experiments show that important social group types can be categorized in an unsupervised manner by using these cues.

5.3 Future Directions

Accurate unsupervised head pose estimation methods

The advance in unsupervised head pose estimation make the approach more peactical for real scenarios. However, there are still a lot of remaining challenges to be solved. One of them is the assumption of the approach itself. In our method, human walking direction is used as a cue to their head poses. However, by studying how human paid attention to their surroundings, we believe there are still vast arrays of cues other than walking direction to be considered.

Another possible direction is to utilize image samples from multiple cameras in order to provide more accurate results. In recent years, scenarios with multiple cameras are becoming more common due to quick advances of camera manufacturing technologies. By the use of multiple cameras, sophisticated measurements can be made in order to acquire ground-truth data in an unsupervised manner and provide accurate results.

Features and estimators for social group detection

Our approach have shown that social groups can be discovered accurately by using attention-based and position-based cues. These cues are originated from analyzing how humans themselves identify social groups from the video. We believe that there are still other cues humans use to discover social groups, and investigating and discovering these cues will be beneficial to the group discovery task. We also believe that more sophisticated framework can be built by carefully observing social group behaviors of humans.

Human action prediction using social group type cues

Recent advances on human action predictions based on social group memberships have shown promising results to be used in real scenarios. With the addition of information on social group types, even more accurate prediction results can be expected. Due to the unsupervised nature of our approach, it can be easily integrated into these approaches to improve the prediction accuracy. In this sense, in addition to their applications in analytical studies, the discovery of social groups and their types can also be applied to real life scenarios.

Bibliography

[1] A. P. Hare, “Handbook of small group research..,” 1976.

[2] M. Sherif and C. W. Sherif, “An outline of social psychology.,” 1956.

[3] H. Tajfel and J. C. Turner, “An integrative theory of intergroup conflict,” *The social psychology of intergroup relations*, vol. 33, no. 47, p. 74, 1979.

[4] J. C. Turner, “Towards a cognitive redefinition of the social group,” *Social identity and intergroup relations*, pp. 15–40, 1982.

[5] M. Andersen and H. Taylor, *Sociology: the essentials*. Cengage Learning, 2012.

[6] C. H. Cooley, *Social organization*. Transaction Publishers, 1956.

[7] K. Yamaguchi, A. Berg, L. Ortiz, and T. Berg, “Who are you with and where are you going?,” in *Proc. 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1345–1352, 2011.

[8] 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)*, pp. 452–465, 2010.

[9] L. Bazzani, M. Cristani, G. Paggetti, D. Tosato, G. Menegaz, and V. Murino, “Analyzing groups: A social signaling perspective,” in *Video Analytics for Business Intelligence* (C. Shan, F. Porikli, T. Xiang,

and S. Gong, eds.), vol. 409 of *Studies in Computational Intelligence*, pp. 271–305, Springer Berlin Heidelberg, 2012.

[10] 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, 2010.

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

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

[13] 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)*, pp. 1886–1893, 2012.

[14] 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)*, pp. 1–12, 2012.

[15] Z. Qin and C. R. Shelton, "Improving multi-target tracking via social grouping," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[16] J. R. Anderson, *Cognitive psychology and its implications.* Macmillan, 2005.

[17] R. Sternberg, *Cognitive psychology.* Cengage Learning, 2008.

[18] E. Spelke, W. Hirst, and U. Neisser, "Skills of divided attention," *Cognition*, vol. 4, no. 3, pp. 215–230, 1976.

[19] S. 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)*, vol. 33, no. 1, pp. 101–116, 2011.

[20] C.-W. Chen, R. Ugarte, C. Wu, and H. Aghajan, “Discovering social interactions in real work environments,” in *Proc. 2011 IEEE International Workshop on Social Behavior Analysis (SBA)*, pp. 933 –938, 2011.

[21] S. Asteriadis, K. Karpouzis, and S. Kollias, “Visual focus of attention in non-calibrated environments using gaze estimation,” *International Journal of Computer Vision*, vol. 107, no. 3, pp. 293–316, 2014.

[22] A. H. Gee and R. Cipolla, “Determining the gaze of faces in images,” *Image and Vision Computing*, vol. 12, pp. 639–647, 1994.

[23] L.-P. Morency, C. M. Christoudias, and T. Darrell, “Recognizing gaze aversion gestures in embodied conversational discourse,” in *Proc. of the 8th International Conference on Multimodal Interfaces (ICMI)*, ICMI ’06, (New York, NY, USA), pp. 287–294, ACM, 2006.

[24] J.-G. Wang and E. Sung, “Em enhancement of 3d head pose estimated by point at infinity,” *Image and Vision Computing*, vol. 25, no. 12, pp. 1864 – 1874, 2007. The age of human computer interaction.

[25] R. Stiefelhagen, “Tracking focus of attention in meetings,” in *Proc. Fourth IEEE International Conference on Multimodal Interfaces.*, pp. 273 – 280, 2002.

[26] 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*, pp. 949 –952, 2006.

[27] S. 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*, pp. 53 –56, 2008.

[28] S. 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*, vol. 39, no. 1, pp. 16 –33, 2009.

[29] N. M. Robertson and I. D. Reid, “Estimating gaze direction from low-resolution faces in video,” in *Proc. 9th European Conference on Computer Vision (ECCV)*, pp. 402–415, 2006.

[30] B. Benfold and I. Reid, “Colour invariant head pose classification in low resolution video,” in *Proc. 19th British Machine Vision Conference (BMVC)*, 2008.

[31] J. Orozco, S. G. Gong, and T. Xiang, “Head pose classification in crowded scenes,” in *Proc. the 20th British Machine Vision Conference (BMVC)*, 2009.

[32] B. Benfold and I. Reid, “Guiding visual surveillance by tracking human attention,” in *Proc. 20th British Machine Vision Conference (BMVC)*, pp. 14.1–14.11, 2009.

[33] A. Schulz, N. Damer, M. Fischer, and R. Stiefelhagen, “Combined head localization and head pose estimation for video-based advanced driver assistance systems,” in *Proc. 33rd Annual Symposium of the German Association for Pattern Recognition (DAGM)*, pp. 51–60, 2011.

[34] A. Schulz and R. Stiefelhagen, “Video-based pedestrian head pose estimation for risk assessment,” in *Proc. 15th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pp. 1771 –1776, 2012.

[35] Home Office Scientific Development Branch, “Imagery library for intelligent detection systems (i-LIDS).” http://homeoffice.gov.uk, 2007.

[36] N. Gourier, J. Maisonnasse, D. Hall, and J. L. Crowley, “Head pose estimation on low resolution images,” in *Proc. First International Evaluation Workshop on Classification of Events, Activities and Relationships (CLEAR 2006)*, pp. 270–280, 2006.

[37] N. M. Robertson, I. D. Reid, and J. M. Brady, “What are you looking at? gaze estimation in medium-scale images,” in *Proc. 16th British Machine Vision Conference (BMVC2005)*, September 2005.

[38] 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)*, pp. 2344 –2351, 2011.

[39] 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)*, pp. 1544–1551, 2012.

[40] I. Chamveha, Y. Sugano, D. Sugimura, T. Siriteerakul, T. Okabe, Y. Sato, and A. Sugimoto, “Head direction estimation from low resolution images with scene adaptation,” *Computer Vision and Image Understanding (CVIU)*, vol. 117, no. 10, pp. 1502 – 1511, 2013.

[41] W. Ge, R. Collins, and B. Ruback, “Automatically detecting the small group structure of a crowd,” in *Proc. 2009 Workshop on Applications of Computer Vision (WACV)*, pp. 1–8, 2009.

[42] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, “Modeling individual and group actions in meetings with layered hmms,” *Multimedia, IEEE Transactions on*, vol. 8, pp. 509–520, June 2006.

[43] P. Dai, H. Di, L. Dong, L. Tao, and G. Xu, "Group interaction analysis in dynamic context," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 39, pp. 34–42, Feb 2009.

[44] N. Vaswani, A. Chowdhury, and R. Chellappa, "Activity recognition using the dynamics of the configuration of interacting objects," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2, pp. II–633–40 vol.2, June 2003.

[45] S. M. Khan and M. Shah, "Detecting group activities using rigidity of formation," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, (New York, NY, USA), pp. 403–406, ACM, 2005.

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

[47] Z. Tang, A. Castrodad, M. Tepper, and G. Sapiro, "Are you imitating me? unsupervised sparse modeling for group activity analysis from a single video," *CoRR*, vol. abs/1208.5451, 2012.

[48] X. Wang, X. Ma, and W. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, pp. 539–555, March 2009.

[49] K. Smith, S. O. Ba, J.-M. Odobez, and D. Gatica-Perez, "Tracking the visual focus of attention for a varying number of wandering people," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 30, no. 7, pp. 1212 –1229, 2008.

[50] E. Murphy-Chutorian and M. M. Trivedi, “Head pose estimation in computer vision: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 31, no. 4, pp. 607 –626, 2009.

[51] A. Puri, H. Kannan, and P. Kalra, “Coarse head pose estimation using image abstraction,” in *Proc. 9th Conference on Computer and Robot Vision (CRV2012).*

[52] D. Huang, M. Storer, F. De la Torre, and H. Bischof, “Supervised local subspace learning for continuous head pose estimation,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR2011)*, pp. 2921–2928, 2011.

[53] G. Fanelli, T. Weise, J. Gall, and L. V. Gool, “Real time head pose estimation from consumer depth cameras,” in *Proc. 33rd Annual Symposium of the German Association for Pattern Recognition (DAGM2011)*, pp. 101–110, 2011.

[54] G. Fanelli, J. Gall, and L. Van Gool, “Real time head pose estimation with random regression forests,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR2011)*, pp. 617–624, 2011.

[55] N. Krahnstoever, M.-C. Chang, and W. Ge, “Gaze and body pose estimation from a distance,” in *Proc. 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS2011)*, pp. 11 –16, September 2011.

[56] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 23, no. 6, pp. 681–685, 2001.

[57] I. Matthews and S. Baker, “Active appearance models revisited,” *International Journal of Computer Vision (IJCV)*, vol. 60, pp. 135–164, 2004.

[58] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[59] V. Prisacariu and I. Reid, “fastHOG - a real-time GPU implementation of HOG,” Tech. Rep. 2310/09, Department of Engineering Science, Oxford University, 2009.

[60] C. Tomasi and T. Kanade, “Detection and tracking of point features,” tech. rep., CMU-CS-91-132, Carnegie Mellon University, 1991.

[61] 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*, vol. 2, pp. 674–679, 1981.

[62] D. H. Douglas and T. K. Peucker, “Algorithms for the reduction of the number of points required to represent a digitized line or its caricature,” *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 10, pp. 315–354, 1973.

[63] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 22, pp. 888 –905, August 2000.

[64] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001.

[65] B. Benfold and I. D. Reid, “Gaze direction dataset.” http://www.robots.ox.ac.uk/˜lav/Research/Projects/2009bbenfold_headpose/ project.html, 2009.

[66] G. Bradski, “The OpenCV Library,” *Dr. Dobb’s Journal of Software Tools*, 2000.

[67] M. Krinidis, N. Nikolaidis, and I. Pitas, “3d head pose estimation using support vector machines and physics-based deformable surfaces,” in *Proc. 9th International Symposium on Signal Processing and Its Applications (ISSPA2007)*, pp. 1 –4, February 2007.

[68] I. Chamveha, Y. Sugano, Y. Sato, and A. Sugimoto, “Social group discovery from surveillance videos: A data-driven approach with attention-based cues,” in *Proc. 24th British Machine Vision Conference (BMVC)*, 2013.

[69] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S.-C. Zhu, “Cost-sensitive top-down / bottom-up inference for multiscale activity recognition,” in *Proc. 12th European Conference on Computer Vision (ECCV)*, 2012.

[70] F. Rohlf, “Single link clustering algorithms,” in *IBM Research Report RC 8569*, p. 33.

[71] 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)*, pp. 1713–1720, 2011.

[72] D. Chakrabarti, R. Kumar, and A. Tomkins, “Evolutionary clustering,” in *Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’06)*, KDD ’06, (New York, NY, USA), pp. 554–560, ACM, 2006.

[73] J. Sivic and A. Zisserman, “Video google: a text retrieval approach to object matching in videos,” in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 1470–1477 vol.2, Oct 2003.

Publication List

Journal Papers

1. Isarun Chamveha, Yusuke Sugano, Daisuke Sugimura, Teera Siriteerakul, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto, "Head direction estimation from low resolution images with scene adaptation", Computer Vision and Image Understanding (CVIU) 117.10 (2013): 1502-1511.

Conference Papers

1. Isarun Chamveha, Yusuke Sugano, Yoichi Sato, and Akihiro Sugimoto, "Social Group Discovery from Surveillance Videos: A Data-Driven Approach with Attention-Based Cues", Proc. of the 24th British Machine Vision Conference (BMVC2013), September 2013.

2. Isarun Chamveha, Yusuke Sugano, Daisuke Sugimura, Teera Siriteerakul, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto, "Appearance-Based Head Pose Estimation with Scene-Specific Adaptation", Proc. of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), November 2011.

Technical Reports

1. Isarun Chamveha, Yusuke Sugano, Daisuke Sugimura, Teera Siriteerakul, Takahiro Okabe, Yoichi Sato and Akihiro Sugimoto, "環境への自

動適応を伴うアピアランスベース頭部姿勢推定", 情報処理学会研究報告コンピュータビジョンとイメージメディア (CVIM), September 2011.