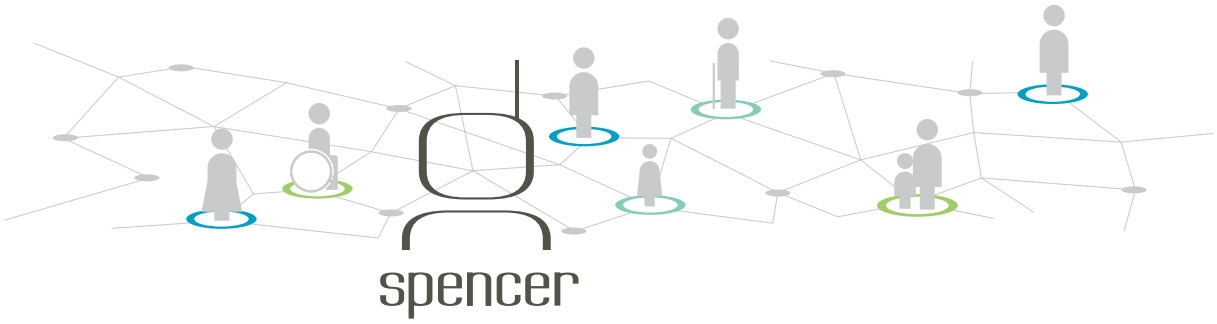

Grant agreement no: FP7-600877

SPENCER:

Social situation-aware perception and action for cognitive robots

Project start: April 1, 2013

Duration: 3 years

DELIVERABLE 4.6

Social relation analysis

Due date: month 34 (May 2013)

Lead contractor organization: UT

Dissemination Level: PUBLIC

Contents

1 Introduction

A key element for understanding social context for robots is the recognition and representation of social relations between humans in its surrounding. In this deliverable, we focus on social relations as displayed non-verbally by groups of people such as families, couples, or friends, and estimate social hierarchy based on the identification of an appropriate spokesperson within a group of acquainted individuals. To this end, we learn inter-person relations by considering the spatio-temporal evolution of relative track trajectories and their attributes, and by moments of direct communication or conversation.

The document is structured as follows: we address the problem of social relation analysis as estimating social hierarchy in Section 2 and briefly describe the implemented ROS module. In Section 3, we frame the problem as predicting probabilities of individuals belonging to the same group from perceptual cues and using those probabilities to extract grouping hypotheses from a weighted social network graph.

2 Social Hierarchy Analysis: Spokesperson Detection (T4.4)

In this section, we present experimental study and discussion on the problem of spokesperson detection as one variant of social relation analysis, then describe the social involvement analysis, and the corresponding ROS module which has been implemented.

2.1 Experimental Study

Task 4.4 involves the spokesperson detection task for the SPENCER robot. Building on the relationship between speaking and group roles, social involvement based on nonverbal cues, experiments on 1) automated speech detection using nonverbal cues from videos (Deliverable D4.5), and 2) social involvement based on nonverbal cues from static images are studied.

2.1.1 Speech Detection

As described in Deliverable D4.5, We found the relationship between gesturing and speaking is not one-to-one. Gesturing indicates speech, but up to half of the speech that was observed in our data is not accompanied by strong gesturing. Furthermore, the relationship between gesturing and speaking is likely person-specific. Overall, the main picture that is painted by all experiments and results combined is that although we are able to detect speech above chance, we have not found a way yet to deal with the speech that is not supported by obvious gesturing. This is partly due to the limited size of the data sets that we were conducting experiments on.

Our experiments have shown that a trade-off can be made between the resolution of the head pose and the number of frames of observations that are required to make an accurate estimate of a conversing group.

Our experiments also showed that the dynamics of interaction in conversations of larger group sizes differed greatly from those of smaller groups. This will have significant consequences when de-

ciding how to distribute the probability of being a spokesperson between the set of people identified to be in a group. That is, the average mutual information in the coordination of behavior between larger groups tends to be lower and therefore behaviors indicating that they are involved in a conversation may be more sparse and weaker in nature, particularly if observations that are of only a very short duration are available. This will also have consequences for the length of time needed to make a confident estimate of a person's membership in a group and also their level of involvement in it. It further implies that a hierarchical approach will be required that makes decisions about the spokesperson based on the length of observations taken so far and the confidence about each person's level of involvement in a group. For extremely short observations, more decisions will have to be made about the spokesperson using very basic behavioral information such as position and height.

2.1.2 Social Involvement

The investigations in the speaker detection task led us to the conclusion that investigating social involvement as a representation of social hierarchy is very important. Moreover, investigating this problem in still images would also be very important in cases where longer observations are not available. Here we summarize our study since the last deliverable on social involvement in [15] (accepted to CVPR 2016), in which we present the first attempt to analyze differing levels of social involvement in free standing conversing groups (or the so-called F-formations) from static images. In addition, we enrich state-of-the-art F-formation (conversational group) modelling by learning a frustum of attention that accounts for the spatial context. That is, F-formation configurations vary with respect to the arrangement of furniture and the non-uniform crowdedness in the space.

Specifically, we approach the novel problem of detecting *associates* of F-formations. F-formations are defined by psychology theory as [4]; as a spatial organization of people gathered for conversation where each member has an equal ability to sense all other members. These so-called *associates* of F-formations are defined by psychologists as people who are attached to an F-formation but do not have the same status as full members (see Figure 1 (a)). These could be people who are in an F-formation but are not fully involved, or people who are standing on the periphery of an F-formation and want to join but is not allowed by the members in it. In particularly, spatially, the people who standing inside the F-formation are the people that we want to avoid labelling as potential spokespersons of a group.

Datasets. The majority of prior works have considered the labelling of conversing group as an objective task, requiring only a single annotator. We carry out extensive experimental validation of our proposed approach by collecting a novel set of multi-annotator labels of involvement on the publicly available Idiap Poster Data; to our knowledge, the only multi-annotator labelled database of free standing conversing groups that is currently available.

Methodology. We detect associates by modeling its social prior with its associated conversational group (F-formation) based on non-verbal cue obtained by top-down surveillance camera, where a set of scale (group size) and orientation invariant features are used to train the social prior. The flowchart of the methodology is shown in Figure 2. Given the position and body orientation on the ground plane of a set of people, a group detector is first applied to find the conversational group location (F-formation will be used in the following sections to indicate conversational groups); social prior

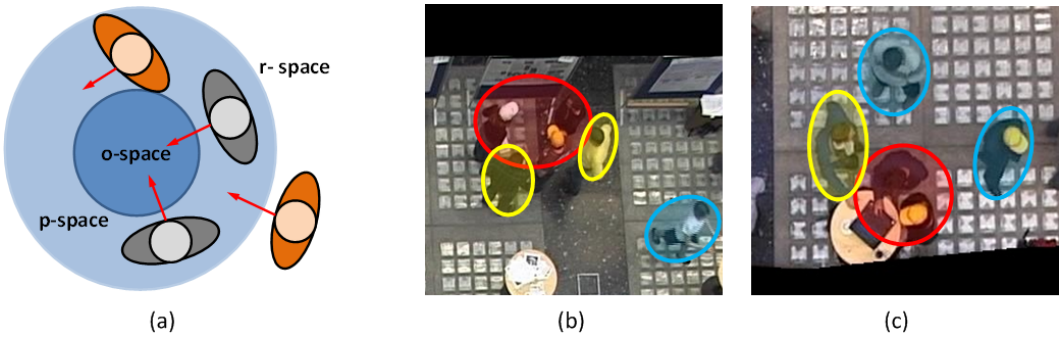

Figure 1: Illustrations of F-formations. (a) The F-formation spaces, gray people stand in the p-space. Red arrows indicate body orientation. Orange people are associates of the F-formation. (b) and (c) example snapshots: F-formations members, associates, and singletons are circled in red, yellow, and blue respectively according to one of our annotators.

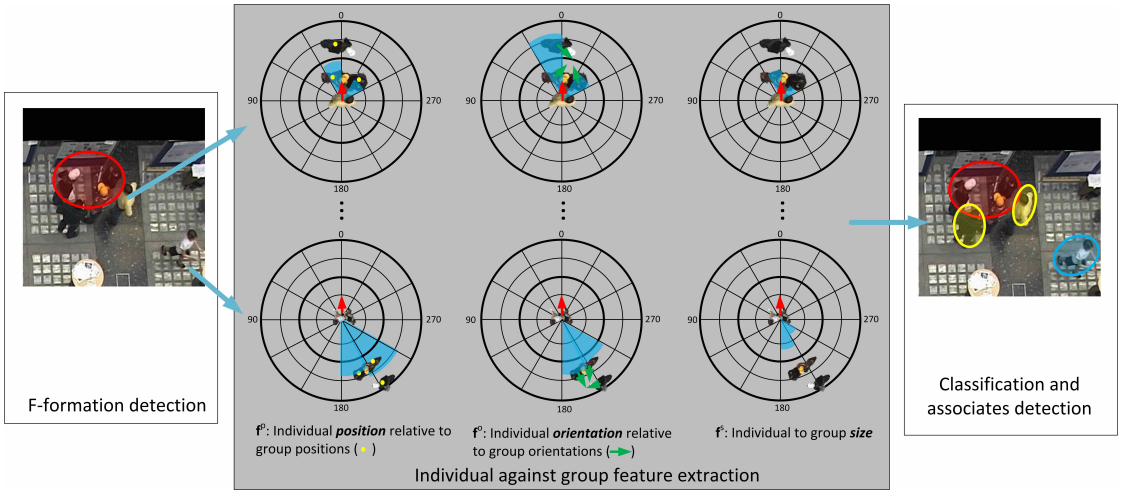

Figure 2: Flow diagram showing the stages of F-formation and associate detection.

features are extracted next from every individual; trained classifiers are then used to determine the involvement of an individual with an F-formation. That is, three labels are generated for each person: full F-formation member, associate of the F-formation, or singleton.

In terms of F-formation detection, we proposed a spatial-context-aware F-formation detector, which models people's frustum of attention in a principled way by taking into account how the people interaction with the spatial layout of furniture and entryways can affect how people stand relative to each other (i.e. the social context) The method is in general more adaptive to different datasets so for example, different frustra of attention parameters can be learned from scenarios with a non-uniform density of crowding.

For associates detection, as shown in the middle image in Figure 2, we proposed a scale/ orientation invariant feature representation to detect associate which accounts for its sparse nature.

Experiment and results. We set up 3 experiments to evaluate our model by measuring precision, recall, and F-measure, the experiments were: 1) F-formation detection 2) associates detection, 3) Improved F-formation detection using feedback from the associates detection. We used state-of-art F-formation detectors **DSFF** [3], **HFF** [2], **ACCVKL** [14], and **ACCVJS** [14] as baselines for experiment 1. Since we are the first to approach the task of detecting associates, we create three baseline detectors to compare with our proposed associate detector. First, **SA** labels all people who are not in an F-formation (mostly singletons) as associates. Second, **RA** labels people as associates of an F-formation if their distance to it is less than or equal to the average distance between pairwise members of F-formations according to the entire labeled data. Third, **ADA** is set based on the average disagreement between annotators where for each pair, we treated one annotation as a detected result to compute performance against another annotation. We also compared performances with different feature combinations (proximity, orientation, and group size).

The results show that, with harsher criterion (full agreement of annotations) our F-formation detector significantly out-performs the state-of-the-art by 10% F-measure with a cross-validated comparison. Using our associate detection model, we were able to discover patterns in proximity and orientation in the behaviours of associates that enable significant improvements over baseline methods with a detection rate of 71% F-measure, which means there are indeed certain patterns of associate behaviour that differs from the behaviour of singletons. By cleaning the detected in-group associates before re-performing F-formation detection, we were able to significantly improve F-formation detection on all cases by 5% F-measure where there was full-agreement amongst annotators on full-members of each F-formation.

The CVPR submission of this work has been amended to the report and is entitled:'Beyond F-formations: Determining Social Involvement in Free Standing Conversing Groups from Static Images'.

2.2 Software Prototype

The earlier study of data collected at Schiphol in year 2 highlighted that the ultimate scenario that we will need for the demonstrator may be very different from the datasets we have been testing on. Our static image based model is good for fast response with such potentially short but dynamic observations.

As a recap, the spokesperson detection task is for understanding group behavior by estimating social hierarchy based on the identification of appropriate spokesperson within a group of acquainted individuals. A spokesperson should play a leading role in the group, *e.g.*, father in a family, senior member in a businessmen group, whom can be recognized from nonverbal cues. Informative nonverbal cues for the spokesperson detection could be based on: *(i)* distance information such as extracted groups from inter-personal nearness cues (T2.3) and relative tracked people's position to the robot (T2.1), *(ii)* attribute information such as gender (T3.1), rough posture (T3.2), head pose (T3.3), and upper-body movement and appearance (T3.4), *(iii)* social information such as social activity detection (T4.2) and social relation analysis (T4.3).

Taking the information above into account we have implemented a basic and fail-safe model. This implements the most robust inference about the social setting and can integrate with the other modules that have been successfully implemented on the robot (mostly distance information obtained

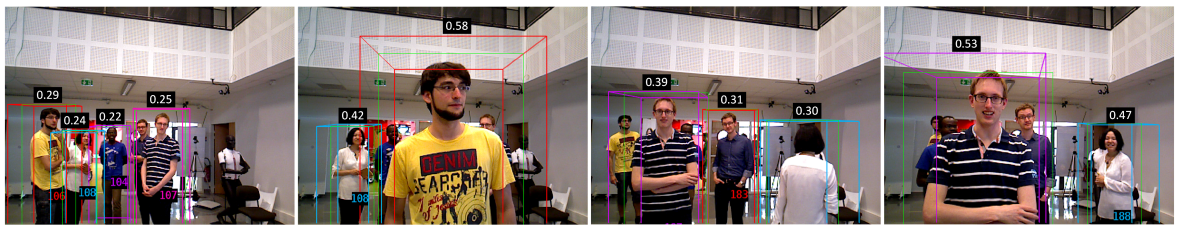

Figure 3: Screen shots of spokesperson probability of detected persons.

from T2.1, and T2.3). We assume that the spokesperson could be the one tending to approach the robot, who might be the closest person to the robot in the group. Therefore, as a software prototype, we have implemented the spokesperson detection module using the following as input:

- Position of the human relative to the robot position (T2.1). For each tracked person published from *spencer_tracking_msgs.msg*, we transform the person's position into the robot-relative base_footprint frame with input from *geometry_msgs.msg*.
- List of detected groups (T2.3).

We output a probability distribution over the tracked people based on their relative position to the robot (the closer to the robot the higher the probability of the spokesperson is) in *spokesperson.msg*:

- `uint32[]` person_id,
- `float64[]` spokesperson_probability

This module is simulated and has been tested with the few short videos we recorded in April, 2015 during the integration week. A few screen shots are shown in Figure 3, where the spokesperson probability of each detection is labeled on top of each pedestrian detection. We can see that the person closer to the robot tends to get higher probability.

3 Social Relation Analysis from High-Level Cues

In addition to the detection of social hierachy and F-formations as examples of analyzing pairwise relations between people from non-verbal cues, this section describes our approach to estimate social relations from motion, activity, and human attribute cues.

For the task of social relation analysis, we seek to estimate probabilities of social relations between in each pair of tracked people. The representation we choose to this end is a fully connected social network graph as used in the contexts of group tracking and spokesperson detection (Fig. 4). In this graph $G = (V, E, A)$ (adopting notation from spokesperson detection), people are represented as nodes V, edges E are social relation candidate between them, and $A = \{ij\}; i, j \in V$ is a probability or affinity function that defines "closeness" between each pair of people.

Computing pairwise social relation probabilities A can be done from different perceptual cues. Concretely, in SPENCER, we have studied the following cues:

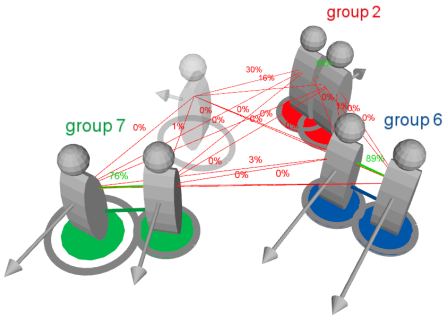

Figure 4: A example social network graph $G = (V, E, A)$ between tracked persons. Individuals are the nodes V, edges E denote social relation candidates between pairs of people weighted by an associated probability A. These probabilities can be determined from motion cues as well as higher-level recognition cues.

- **Coherent motion indicators (Task T2.3):** Coherent motion indicators are features derived from the movements of surrounding people. Empirical studies in social science have found three dominant motion indicators for people that walk in groups [11]: relative distance, relative orientation, and similar velocity to their direct neighbors. Fig. 4 illustrates an example in which the edges of the social network graph are labelled with probabilities computed in this way. The tracking system developed in task T2.3 uses a probabilistic SVM classifier to predict these probabilities for tracking groups of people.

- **Social activity detections (Task T4.2):** Being engaged into the same activity is a piece of information for detecting groups and thereby analyzing social relations between individuals. In this view, activity recognition is a form of social relation analysis, pursued in SPENCER in task T4.2. Therein, we define social activities to be functions of joint interpersonal movements and activities of groups of people, their individual attributes and social relations. We have developed a object-centric feature representation based on shape for classification and use an on-line variant of Conditional Random Fields for smoothing in particular to handle missing attribute information due to occlusions of people. With this system (see [12] for first results), we are able to recognize five different activity types: queueing, moving with flow, standing, moving against flow and shopping (for details see Deliverable D4.4 with attached paper under review).

- **Human attributes (Task T3.1):** Knowing more about individuals in the form of attributes such as age, gender or clothing is a meaningful cue to infer social relations between tracked people. Similar attributes such as clothing or dissimilar attributes such as gender (e.g. to predict couple-like social relations between male and female individuals) are strong patterns to infer more complex social relations. This has been done in task T3.1 (Deliverable D3.1 and [7, 5]) using a boosted tessellation approach which outperformed two deep-learning baselines in our experiments. Concretely, we are able to recognize the following attributes using RGB-D data: gender, has long hair, has long trousers, has long sleeves and has jacket with accuracies up to 90% in real-time (see [7, 5]).

- **Human posture cues (Tasks T3.2, T3.3, T3.4):** Similar to human attributes, human posture cues such as upper body posture (T3.4), head orientation (T3.3) or rough full body pose (T3.3)

are highly informative cues for the task of social relation analysis. The recognition of those cues enables a robot to detect groups of people that stand together facing each other with their body and head (for instance in F-formations as studied above for spokesperson detection) or walk together (same body orientation) looking at each other with difffernt head orientations. See Deliverable D3.3 for details on the recognition systems developed in tasks T3.2, T3.3, and T3.4.

Research on these perceptual cues for social relation analysis in SPENCER was highly successful and led to publications [10, 8, 12, 9, 5, 7, 1, 6, 15] with several of them collaborations between consortium members (in particular ALU-FR with TUM and ALU-FR with RWTH) and several more submissions under review.

In addition to the extraction of such cues from sensory data, the task of social relation analysis includes the estimation of social grouping hypothesis from pairwise relation probabilites A as weights to the edges of the graph. There are several ways to obtain a social grouping hypothesis from a weighted graph. In SPENCER, we have developed two different approaches: graph-cut algorithms followed by a multi-model social grouping hypothesis tracker as used in task T2.3 [10, 8]. This approach performs multi-hypothesis tracking of social grouping models in real-time and is able to recover from wrong grouping hypotheses when more information accumulates evidence that a different than the current-best hypothesis is more probable. The second approach is modeling social groupings as dominant sets which can be detected using quadratic programming techniques as shown in the spokesperson detection task (see above). This approach relies in turn on the method by Pavan and Pelillo [13].

These two steps, extraction of non-verbal cues from sensory data and estimation of social relations between individuals from such cues constitute the social relation analysis task as defined in task T4.4 in the DoW.

4 Conclusions

For Tasks T4.3 and T4.4 we addressed the problem of social relation analysis from non-verbal cues using different approaches. While task T4.3 aimed at high-level perceptual cues and the extraction of grouping information from a social network graph, task T4.4 aimed at the detection problem of social hierachy and F-formations. For both tasks, we proposed different spatial-context-aware feature representations as well as novel detectors such as the one for F-formations which models people's frustum of attention in a principled way while considering the influence of the social and spatial context. The developed methods in SPENCER extend the state-of-the-art as reflected by the high number of peer-reviewed publications, also when for training the model parameters not much data was available.

References

[1] S. Breuers, S. Yang, M. Mathias, and B. Leibe. Exploring bounding box context for multi-object tracker fusion. In *IEEE Winter Conference on Applications of Computer Vision (WACV), 2016*, 2016.

[2] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino. Social interaction discovery by statistical analysis of F-formations. In *BMVC*, pages 1–12, 2011.

[3] Hayley Hung and Ben Kröse. Detecting f-formations as dominant sets. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 231–238. ACM, 2011.

[4] Adam Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990.

[5] T. Linder and K. O. Arras. Real-time full-body human attribute classification in rgb-d using a tessellation boosting approach. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ Int. Conf. on*, 2015.

[6] T. Linder, S. Breuers, B. Leibe, and K. O. Arras. Taking a closer look at people tracking in challenging environments using a novel multi-modal evaluation framework. In *Robotics and Automation (ICRA), 2016 IEEE Int. Conf. on*, 2016.

[7] T. Linder, S. Wehner, and K. O. Arras. Real-time full-body human gender recognition in (rgb)-d data. In *Robotics and Automation (ICRA), 2015 IEEE Int. Conf. on*, 2015.

[8] Timm Linder and Kai O. Arras. Multi-model hypothesis tracking of groups of people in RGB-D data. In *IEEE Int. Conf. on Information Fusion (FUSION'14)*, Salamanca, Spain, 2014.

[9] Timm Linder, Fabian Girrbach, and Kai O. Arras. Towards a robust people tracking framework for service robots in crowded, dynamic environments. In *Assistance and Service Robotics Workshop (ASROB-15) at the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS) 2015*, Hamburg, Germany, 2015.

[10] Matthias Luber and Kai O. Arras. Multi-hypothesis social grouping and tracking for mobile robots. In *Robotics: Science and Systems (RSS'13)*, Berlin, Germany, 2013.

[11] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. *PLoS ONE*, 5(4), April 2010.

[12] Billy Okal and Kai O. Arras. Towards group-level social activity recognition for mobile robots. In *IROS 2014 Workshop on Assistance and Service Robotics in a Human Environment*, Chicago, USA, 2014.

[13] Massimiliano Pavan and Marcello Pelillo. Dominant sets and pairwise clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):167–172, 2007.

[14] Sebastiano Vascon, Eyasu Zemene Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo, and Vittorio Murino. A Game-Theoretic Probabilistic Approach for Detecting Conversational Groups. In *ACCV*, 2014.

[15] Lu Zhang and Hayley Hung. Beyond F-formations: Determining Social Involvement in Free Standing Conversing Groups from Static Images (submitted). In *conference on Computer Vision and Pattern Recognition*, 2016.

Beyond F-formations: Determining Social Involvement in Free Standing Conversing Groups from Static Images

Lu Zhang[1,2] and Hayley Hung[1]

[1]Delft University of Technology, Mekelweg 2, Delft, Netherlands,
{lu.zhang,h.hung}@tudelft.nl
[2]University of Twente, Drienerlolaan 5, Enschede, Netherlands

Abstract

In this paper, we present the first attempt to analyse differing levels of social involvement in free standing conversing groups (or the so-called F-formations) from static images. In addition, we enrich state-of-the-art F-formation modelling by learning a frustum of attention that accounts for the spatial context. That is, F-formation configurations vary with respect to the arrangement of furniture and the non-uniform crowdedness in the space during mingling scenarios. The majority of prior works have considered the labelling of conversing group as an objective task, requiring only a single annotator. However, we show that by embracing the subjectivity of social involvement, we not only generate a richer model of the social interactions in a scene but also significantly improve F-formation detection. We carry out extensive experimental validation of our proposed approach by collecting a novel set of multi-annotator labels of involvement on the publicly available Idiap Poster Data; to our knowledge, the only multi-annotator labelled database of free standing conversing groups that is currently available.

1. Introduction

In recent years, the analysis of mingling scenarios has received growing attention. The potential of studying social patterns of behaviour in visual scenes has great potential with the recent advances in social signal processing [21]. Potential applications include enabling robots to approach a group and offer assistance in a socially intelligent manner [18], or social surveillance [3], image interpretation or retreival [14].

A major challenge in visual scene interpretation is addressing the problem of bridging the semantic gap [14], which defines the disconnect between information that can be extracted from the pixels in an image and how a human might interpret its contents. Traditionally, this gap has been attributed to the mapping of imagery data to objective interpretations such as the labelling of objects or activities in a scene. However, in recent years, scene analysis has started to consider more complex and subjective concepts such as safety [11] or ambiance [12]. Similarly, in the area of social surveillance [3], researchers have been trying to ascribe social meaning to social scenes. However, unlike conventional scene analysis, social surveillance bridges a more complex semantic gap that associates observable behavioural cues to social phenomena. We call this the *social semantic gap*. Since social phenomena are extremely complex, it is desirable to use findings from social psychology to help inform how visually observed behaviours could be linked to social phenomena to help bridge the gap in an informed manner.

Given the great advances already in person tracking and orientation detection, we focus on how these solutions can be used as behavioural input for bridging the social semantic gap. Specifically, we approach the novel problem of detecting *associates* of conversing groups (or the so-called F-formations). F-formation are defined by psychology theory as [8]; as a spatial organization of people gathered for conversation where each member has an equal ability to sense all other members. These so-called *associates* of F-formations are defined by psychologists as people who are attached to an F-formation but do not have the same status as full members (see Figure 1 (a)).

To the best of our knowledge, state-of-the-art methods for F-formation detection [6, 2, 13, 19, 20] have made three simplifying assumptions. First, each individual is assumed to have a binary membership to an F-formation and to our knowledge, no work has considered refining and enriching this model to label individuals who are partially involved in it. Second, global parameters for the frustrum of attention of each person have been used for the entire visual scene.

However, psychology theory has cited the relaxation of the geometric model of an F-formation when considering the spatial constraints of a room and the furniture in it [8]. Finally, aside from Hung et al. [6], we believe that no other works have seriously addressed the inherently subjective nature of F-formation detection. Our experiments show that by considering the inherent subjectivity of the task, we are better able to model the social scene. That is, by performing associate detection, we show that we can also significantly improve performance on the F-formation detection task.

Concretely, we offer the following contributions; First, we address the novel task of detecting associates of F-formations and propose a novel feature representation that copes with learning from sparse training data. We also show that the state-of-the-art model for full members of F-formations [19] are not appropriate for the modelling of associate behaviour. Second, we model the spatial context of a scene for better F-formation and associate detection by learning a location-dependent frustum of attention of individuals in the scene. Moreover, we address the problem of learning the relative weighting between proximity and orientation given the spatial context of furniture. Third, we contribute new multi-annotator labels on the publicly available Idiap Poster Dataset [6] for modeling associates. Finally, we carry out a deep evaluation and analysis of associates to investigate the complexity of this novel task.

2. Definitions

F-formations and their Associates The psychologist Kendon [8] defined a single conversing group as an F-formation; as a spatial and orientational organization of individuals where each member has equal access to all other members of the group. An F-formation usually consists of three parts, see Figure 1 (a). The o-space is a convex empty space surrounded by the F-formation members, in which every participant orientates themselves inwards, and no external people are allowed. The participants themselves stand in the p-space, which is a narrow strip surrounding the o-space, while the area beyond is called the r-space. Its definition has made it a popular detection task as it relates well to finding maximal cliques in edge-weighted graphs [6, 19, 20].

In practice, a geometric model of a conversing group should be adapted when considering the spatial constraints of a room and the furniture in it [8]. For instance, people talking in front of a laptop may stand closer and look at the same direction (see Figure 1(c) still maintains an F-formation although their o-space could be violated.

Unlike full members of F-formations, Kendon [8] defines associates to be people who are attached to an F-formation but who are not fully involved in the conversation. Associates can be people who try to join an F-formation but are not fully accepted by the group, or can leave an F-formation abruptly without disturbing the conversation. We name these out-group and in-group associates respectively as the former tends to stand in the r-space while the latter tends to stand in the p-space. Another example of an associate could be someone who is waiting for a full member (e.g. their spouse) to leave the F-formation and is not interested in engaging in the conversation.

While F-formations can easily be modelled by either maximal cliques [6, 19, 20] or a joint centre-of-focus in the o-space [2], associate behaviours are not so clearly linked to a single set of social cues. Therefore, the associate detection problem bridges a wider gap and the nature of the problem and how to solve it cannot be so easily translated into a single set of geometric constraints. From the perspective of semantic labelling of a scene, we must also consider that distinguishing full members of F-formations from associates and also singletons is quite important conceptually. Singletons have no social influence on the groups around them. Full F-formation members have the most potential to influence on other members of the groups. Meanwhile, associates have the least potential to influence full members but could be influenced by them. Moreover, in-group associates could be mistaken for full F-formation members and out-group associates for singletons.

Frustum of Attention The frustum of attention [19] (or transactional segment, as defined by Kendon [8] can be considered as a cone-like region extending from the body that represents the spatial and angular extent at which someone is able to see, hear, and potentially touch something or someone else. It represents a three-dimensional space around the human body in which most of our senses and actions are able to be deployed for social interaction. Prior studies have shown that head pose [15, 16, 19], body pose [6], gaze [16, 7], and proximity [6] often provide reliable features for F-formation modeling.

Recent state-of-the-art approaches have tended to use sampling methods to approximate the frustum of attention where the parameters are set carefully by grid search on the entire dataset and the same global model for the frustum of attention is used [19, 2, 13]. There are two main drawbacks of this approach. First, the parameters are likely to overfit on a certain dataset due to the same data being used for training and testing. Second, the variation in F-formation shape caused by the furniture arrangement and non-uniform densities in the crowding of the scene cannot be captured. For example, people can tend to crowd more densely around the area of a bar area even if they are not trying to order drinks or lean on it.

3. Related Work

Exploiting the frustum of attention is very important for detecting F-formations, studies have showed that head pose

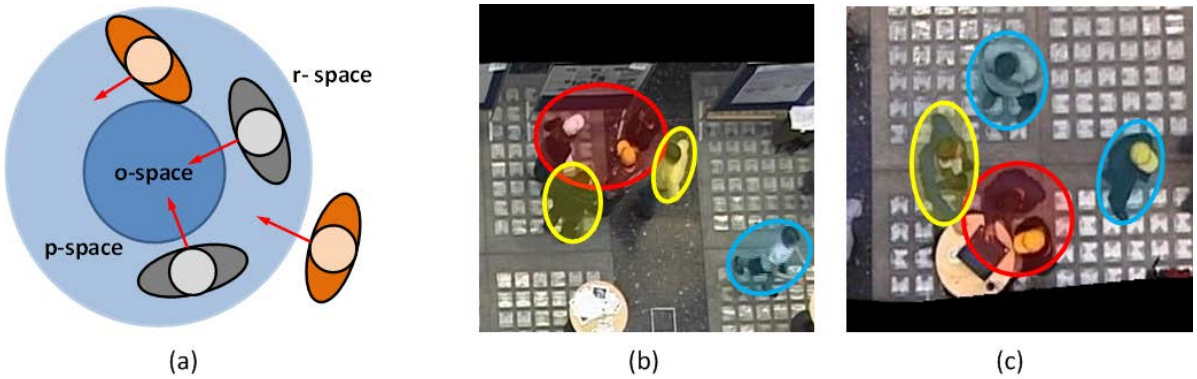

Figure 1. Illustrations of F-formations. (a) The F-formation spaces, gray people stand in the p-space. Red arrows indicate body orientation. Orange people are associates of the F-formation. (b) and (c) example snapshots: F-formations members, associates, and singletons are circled in red, yellow, and blue respectively according to one of our annotators.

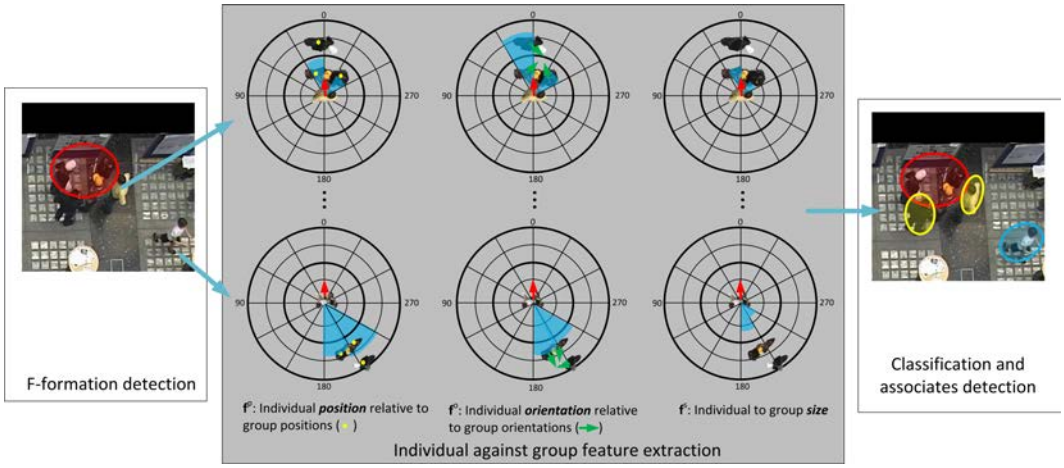

Figure 2. Flow diagram showing the stages of F-formation and associate detection.

[15, 16, 19], body pose [5], gaze [16, 7], and proximity [6] often provide reliable patterns. In [22], F-formations are detected by estimating people's position and lower body orientation using only their head position and orientation from a single camera. The modularity cut algorithm [9] was proposed to identify F-formations from automatically extracted trajectories by [23]. To our knowledge, in terms of the treatment of hierarchy in groups, the work of [23]. is quite close to ours as they proposed to used eigendecomposition to find centrality in a large mingling group of people. Unfortunately, the data they used was staged but showed participants with high centrality to be those who mingled with more different people.

A Hough voting strategy was proposed in [2], which estimates the location of o-space by density estimation. The size of F-formation is modeled based on Hough voting strategy in [13]. In [6, 19], detecting F-formations is considered as a clustering problem, where each person is defined as a node in the graph, and each edge is the "closeness" between a pair of people. The goal is to find a dominant set [10] in the graph and the edges of the graph is computed based on body orientation and proximity. In [19], the temporal information is added in the dominant set based approach. A density-based approach has been proposed in [4] where the final purpose of the task was to dynamically select camera angles for automated event recording. In [17], temporal patterns of activities have been subsequently analyzed. In this paper, we will follow the dominant set framework because it gives reliably good results in general [19] and enables a systematic explanation of the learned model so we can interpret better the social phenomena at play in the experimental data. In contrast to the growing numbers of works on F-formation detection, to our knowledge, no one has attempted to detect associates before.

4. Data

We used the publicly available Idiap Poster Data [6] [1], which consists of 3 hours of aerial video of over 50 people

[1] https://www.idiap.ch/dataset/idiap-poster-data

during a scientific poster session and coffee break. In this poster session, posters are put around the perimeter of the scene, two small round tables are located in the middle and bottom of the image, a drinks table is located in the bottom right of the image, two entrances are located at the far left and top right of the scene. A screen shot is shown in the left of Figure 4. In total, 82 images including 1700 instances of people were annotated by 24 paid annotators, where each image was annotated by 3 annotators. No consecutively selected images contained the same set of formations. We used the positions and body orientation provided separately by Hung et al. [6]. We augmented this data by adding annotations of associates of the F-formations.

We analyzed the annotations to see whether there was full agreement between the annotators about all members of an F-formation and associate people. 211 instances of associates were annotated. 84 associates were identified with majority agreement (39.8%) and 34 for full agreement (16%). We computed the F1 score considering one annotation as ground truth and one other annotation as detection for each set of data annotated by the same 3 annotators. The mean and standard deviation of the F1 score are 44% and 13% respectively, which shows that associates are not as straight forward to label compared to F-formations (94.74% mean average F-measure when computing the agreement for F-formations from the data). We consider all the annotated associates have different levels of less involvement to groups, that can be visualized by annotators.

To explore the relative angle and orientation relationship between different types of associates of F-formations, we computed histograms of both relative orientation differences between an associate and also distance to closest their nearest F-formation member as shown in the top and bottom parts of Figure 4(b) respectively. The relative orientation of associates to their closest F-formation member has a large peak in probability mass at 0, and $\pi/3$ while there is only a single peak in the bottom histogram, showing that associates tended to stand similarly closely to their nearest F-formation member. The double peak seen in the relative orientation aligns with the idea of associates who are standing in the p-space of an F-formation but appear less involved in the conversation (in-group associates) and those that stand in the r-space, facing towards the F-formation (out-group associates).

5. Methodology

In this paper, we detect associates by modeling its social prior with its associated conversational group (F-formation) based on non-verbal cue obtained by top-down surveillance camera, where a set of scale (group size) and orientation invariance features are used to train the social prior. The flowchart of the methodology is shown in Figure 2. Given the position and body orientation on the group plane of a set of people, a group detector is first applied to find the conversational groups location (F-formation will be used in the following sections to indicate conversational groups); social prior features are extracted next from every individual people; trained classifiers will be used to determine the involvement of a certain people to a F-formation, for instance, F-formation member, associate, or singleton. The modules are described in the following subsections separately.

5.1. Modeling the F-formation as a Dominant set

Building on prior work [6, 19], we exploit the dominant set framework. In an image, people can be represented as a graph $G = (V, E, A)$, where the nodes V are people, E is the set of connections between people, and $A = \{a_{ij}\}\,,\ i, j \in V$ is affinity function defines "closeness" between each pair of people. Given a subset S of the set of of nodes in the graph, the average weighted degree of a node $i \in S$ with respect to set S is $k_S(i) = \frac{1}{|S|}\sum_{j\in S, j\neq i} a_{ij}$. The relative affinity between node $j \notin S$ and i is $\phi_S(i, j) = a_{ij} - k_S(i)$, and the weight of each i with respect to a set $S = R \cup \{i\}$ is defined as

$$w_S(i) = \begin{cases} 1 & |S|=1 \\ \sum_{j\in R} \phi_R(j, i) w_R(j) & otherwise\ , \end{cases} \quad (1)$$

which measures the overall relative affinity between i and the rest of the nodes in S. The relationship between internal and external nodes of a dominant set S is defined as

$$w_S(i) > 0,\ \ \forall i \in S \quad (2)$$

$$w_{S\cup\{i\}}(i) < 0,\ \ \forall i \notin S. \quad (3)$$

Detecting a dominant set is identical to solving the following standard quadratic programme $\max_{\mathbf{x}} \mathbf{x}^T A\mathbf{x}$, $s.t.\ \mathbf{x} \in \Delta$, where $\Delta = \left\{\mathbf{x} \in R^{|V|} : \sum_{i\in V} x_i = 1,\ x_i \geq 0,\ i = 1, \cdots, |V|\right\}$. The indexes of non-zero x_i are the same as the people indexes of a F-formation, in such a way that a F-formation can be identified. This optimization problem can be solved with a method from evolutionary game theory, called replicator dynamics. The first-order replicator can be represented as $x_i = x_i \frac{(A\mathbf{x})_i}{\mathbf{x}^T A\mathbf{x}}$. Once $\mathbf{x}$ converges, one set of F-formation members are detected. A peeling method is used where the detected group is removed and the replicator dynamics is repeated to find the next F-formation. This peeling method is repeated until the minimum distance of pairwise F-formation members is larger than the maximum distance of detected pairwise F-formation members for a given image. For further details of the method used, see [6, 10].

5.2. Social involvement features

As described in Section 1 associates have a complex behaviour that is strongly related to the F-formation that they

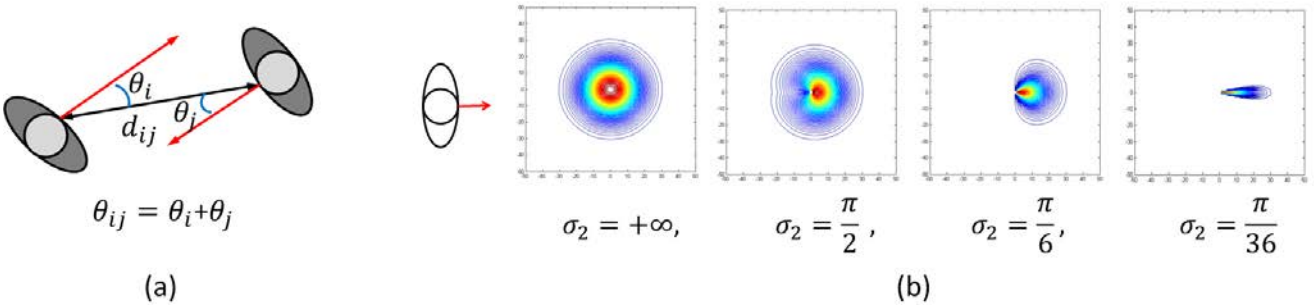

Figure 3. Frustum of attention modeling with body orientation and proximity. (a) Calculation of relative orientation and proximity, (b) frustum of attention map with different parameters. The smaller the σ_2 is, the narrower of frustum attention of a person is.

are associated with. They can exist in either the p-space or r-space. Moreover, unlike the maximal clique constraint of full members of F-formations, associates should be mathematically defined with respect to the spatial arrangement of a candidate set of full members of an F-formation. Searching the space of all possible solutions for the associate and F-formation task is NP. Fortunately, in practice, associates tend to be scatted sparsely enough amongst a set of F-formations so that the maximal clique assumption is not severely disrupted by their presence. Therefore in the first instance, using any existing F-formation detection method to reduce the space of possible hypothesis associate and F-formation pairs is reasonable.

Despite this simplification, another challenge still remains. Due to its sparsity, it is unlikely that a sufficient set of examples exist to account for all possible spatial configurations of an associate and F-formation. Therefore, applying similar features that were used to define full members will lead to a representation that is too sparse to learn from. To make sufficiently descriptive features, we hypothesise therefore that they must be both invariant to the rotation of the associate relative to the group, and also insensitive to the size of the group.

To better understand associates and avoid wrong F-formation detection in the earlier step (*e.g.*, detecting associates as F-formation members), every individual person in the experiment is considered as a associate candidate, so that an associate candidate can be a F-formation member, an associate, or a singleton in practice. Three sets of social prior features $\mathbf{f} = [\mathbf{f}^p, \mathbf{f}^o, \mathbf{f}^s]$, centered at the associate candidate, are extracted to represent the geometric relationship of an associate candidate and its associated F-formation, where the features are based on proximity, body orientation, and group size, respectively. The closest F-formation C to a certain associate candidate $\mathbf{p}_a$ is considered as the associated F-formation of this associate candidate, and $\mathbf{p}_k$ indicates the location of k^{th} F-formation member in C. Note that, $\mathbf{p}_a$ will be removed from C if it is detected as a F-formation member in the earlier F-formation detection step.

Each set of social prior feature $\mathbf{f}$ is a 12-bin histogram, which is defined based on the angle of the vector between F-formation member $\mathbf{p}_k$ and associate candidate $\angle(\mathbf{p}_k - \mathbf{p}_a)$, so that every bin covers an angle of $\pi/6$. We define the m^{th} bin of the three sets of features as

$$\mathbf{f}_m^p = \frac{1}{Z_d \cdot |C_m|} \sum_{k \in C_j} \|\mathbf{p}_k - \mathbf{p}_a\|, \tag{4}$$

$$\mathbf{f}_m^o = \frac{1}{Z_o \cdot |C_m|} \sum_{k \in C_j} (\angle\mathbf{p}_k - \angle\mathbf{p}_a), \tag{5}$$

$$\mathbf{f}_m^s = \frac{1}{Z_s} |C_m|, \tag{6}$$

where the set of F-formation members located in this bin is C_m. We use $\mathbf{f}_m^p$ to represent the average proximity between F-formation members in C_m and $\mathbf{p}_a$, $\mathbf{f}_m^o$ to represent the average relative body orientation between F-formation members in C_m and $\mathbf{p}_a$, and $\mathbf{f}_m^s$ to represent the relative people density in C_m. The features are normalized by Z_d, Z_o, and Z_s, where Z_d is the maximum proximity between associated F-formation members and associate candidate, $Z_o = 2\pi$, and Z_s is the maximum F-formation size. The middle image in Figure 2 shows examples of scale/ orientation invariance feature representation of an associate and a singleton, which encode people's relative location, orientation and group size in. Associates Detection is challenging because they are likely to be detected as F-formation members compare to singletons who are usually far away from F-formation. We use one-vs-the rest strategy to train an associates detector. In the experiment, we compare a set of classifiers, such as Parzen, RBF SVM, Random Forests, and AdaBoost, with 10 fold cross validation. Parzen classifier gives the best performance on our dataset. In our experiment, we use 211 instances of all the annotated associates, 235 fully agreement of singletons and 450 fully agreement of F-formations as training data.

5.3. Training Affinity matrix

To detect F-formations in a complex environment, we need to model the variation of the density of geometric variations of potential F-formations in the space. To capture

this variation, the affinity matrix $\mathbf{A}$ is key. In this paper, we only consider the proximity and body orientation. The "closeness" between people i and j is defined as

$$a_{ij} = e^{-\frac{d_{ij}^2}{\sigma_1^2} - \frac{\theta_{ij}^2}{\sigma_2^2}}, \tag{7}$$

where d_{ij} is the Euclidean distance between two people, θ_{ij} is the sum of difference between each body orientation and the angle of the vector between two people (see Figure 3 for details), and σ_1 and σ_2 are the parameters to be learned. As the values of σ_1 and σ_2 decrease, a person is likely to stand closer and angle more directly towards the others in the F-formation (see Figure 4 (a)). Likewise, as σ_1 and σ_2 increase, members of an F-formation will tend to stand further apart and orientate themselves less directly towards others (see Figure 4 (a)). The objective function is defined as

$$\ell = \sum_{n=1}^{N} 1 - \frac{C^{\{n\}} \cap \hat{C}^{\{n\}}}{C^{\{n\}} \cup \hat{C}^{\{n\}}} \tag{8}$$

where n is the index of an F-formation in an image, N is the total number of annotated F-formations, and $C^{\{n\}}$ and $\hat{C}^{\{n\}}$ are the nth detected set of F-formation members and its corresponding annotation respectively. We consider a detection C and an annotation $\hat{C}$ to match with each other if $\frac{|C \cup \hat{C}|}{|C \cap \hat{C}|} \geq \frac{2}{3}$. Considering that the shape of the F-formation can be influenced by the furniture arrangement, we learn parameters σ_1 and σ_2 as a function of a person's location $\mathbf{p}$. We perform the parameter update in a passive-aggressive way [1]; we only update once per person when the detection goes wrong.

$$\sigma_s(\mathbf{p}) = \sigma_s(\mathbf{p}) - g_s(C)\Delta\sigma_s, \;\; s \in \{1, 2\}. \tag{9}$$

Here, $\Delta\sigma_s$ is the basic step size, which is set to a small value. An adaptive parameter g helps to adapt to different F-formation geometric variations. Given F-formation C, the adaptive parameter g is defined as

$$g_1(C) = y\frac{\|\sum_{i,j\in\hat{C}^{\{n\}}} \hat{d}_{ij} - \sum_{i,j\in C^{\{n\}}} d_{ij}\|}{\sum_{i,j\in\hat{C}^{\{n\}}} \hat{d}_{ij}}, \tag{10}$$

$$g_2(C) = y\frac{\|\sum_{i,j\in\hat{C}^{\{n\}}} \hat{\theta}_{ij} - \sum_{i,j\in C^{\{n\}}} \theta_{ij}\|}{\sum_{i,j\in\hat{C}^{\{n\}}} \hat{\theta}_{ij}}, \tag{11}$$

where $y \in \{-1, 1\}$, $y = 1$ indicates a false negative F-formation member in C, while $y = -1$ indicates false positive F-formation member. Here $\hat{d}$ and $\hat{\sigma}$ are the manually annotated proximity and frustum of attention. In each iteration, we update every person's location in the F-formation.

6. Experiment

6.1. Experiment setup

In the experiment, we initialized $\sigma_1 = 40, \sigma_2 = 30$ for training, whose basic update step sizes were set to $\Delta\sigma_1 = 0.1$ and $\Delta\sigma_2 = \pi/720$ respectively. The numbers of iteration of training for detecting F-formation and associates were both set to 300. Considering that the training samples in each precise location were not distributed densely over the images , we divided the images into blocks of 45×45 pixels where all people located in the same block shared the same learned parameters. We trained using each of the 3 annotations separately, applying 10 fold cross validation for each. Finally, the position and body orientations used to train our models came from the annotations of the Idiap poster data provided by Hung et al. [6].

For evaluation, we consider a group as correctly estimated if at least $(T \cdot |C|)$ of their members are detected, where $|C|$ is the cardinality of the labeled group C, and $T \in [0, 1]$ is an arbitrary threshold; in [2], the scoring threshold $T = 2/3$, corresponds to finding at least two thirds of the members of a group. Here we also consider $T = 1$, to mean that a group is correctly detected only if all members are labeled correctly. From these metrics we calculate standard precision, recall and F1 measures in each frame, averaging them over all the frames and the three sets of annotations. Associates are evaluated by calculating precision, recall and F1 score in the same way, where only the harder $T = 1$ criterion for success is used.

Here, a baseline detector global-**F** is added, which only uses the initialized training value $\sigma_1 = 40, \sigma_2 = 30$ for detecting F-formation. We also compared the performance of our spatial-aware F-formation detector (Spatial-**F**) with state-of-the-art **DSFF** [6], **HFF** [2], **ACCVKL** [19], and **ACCVJS** [19].

Since we are the first to approach the task of detecting associates, we create three baseline detectors to compare with our proposed associate detector (social-**A**). Each baseline result was generated using the annotated data and not detections. First, **SA** labels all people who are not in an F-formation (mostly singletons) as associates. Second, **RA** labels people as associates of an F-formation if their distance to it is less than or equal to the average distance between pairwise members of F-formations according to the entire labeled data. Third, **ADA** is set based on the average disagreement between annotators where for each pair, we treated one annotation as a detected result to compute performance against another annotation. We also compared performances with different feature combinations (**p** is proximity features, **o** indicates orientation features, and **s** is group size features). The associates detector global-**A** extract features based on global-**F** F-formation detection.

We also studied how associates detection can help up im-

Table 1. F-formation detection results with soft ($T = 2/3$) and hard ($T = 1$) criteria for deciding on whether an F-formation is correctly detected.

Method	T=2/3			T=1		
	Prec.	Rec.	F1	Prec.	Rec.	F1
DSFF [6]	0.93	0.92	0.92	0.81	0.81	0.81
HFF [2]	0.93	0.96	0.94	0.81	0.84	0.83
ACCVKL [19]	0.90	0.94	0.92	-	-	-
ACCVJS [19]	0.92	0.96	0.94	-	-	-
global-F	0.87	0.92	0.89	0.72	0.76	0.74
spatial-F	0.91	0.98	0.94	0.91	0.98	0.94

Table 2. Associate detection results. **SA**: labels all singletons as associates, **RA:** labels people close to F-formation as associates, **UA**: performance based on annotator disagreement, global-**A**: use global-**F** detector to extract features, and social-**A**: our proposed detector (details in Sec. 6.1).

Method	Prec.	Rec.	F1
SA	0.06	1.00	0.11
RA	0.11	0.84	0.19
ADA	0.44	0.44	0.44
global-A(p+o+s)	0.89	0.59	0.71
social-A(p)	0.87	0.58	0.69
social-A(o)	0.91	0.55	0.69
social-A(s)	0.78	0.53	0.63
social-A(p+o)	0.89	0.57	0.70
social-A(p+s)	0.85	0.56	0.67
social-A(o+s)	0.91	0.56	0.69
social-A(p+o+s)	0.89	0.59	0.71

Table 3. F-formation detection with associate detection feedback, results are evaluated only on full-agreement annotated F-formations. FB-global-F and FB-spatial-F are detectors with associate detection feedback (details in Sec. 6.1).

Method	Prec.	Rec.	F1
global-F	0.75	0.94	0.83
FB-global-F	0.82	0.94	0.88
spatial-F	0.76	1.00	0.86
FB-spatial-F	0.84	1.00	0.91

prove F-formation detection. As the F-formation detector has problem mostly with in-group associates, so that using detected associates to clean up detected false positive F-formation is helpful. We remove F-formation members detected as associates. The performances of Spatial-**F** and global-**F** are evaluated with $T = 1$ hard criterion and full agreed annotated F-formations since there is no associate labels from any annotator.

6.2. F-formation Detection Results

Examples of the learned values for σ_1 and σ_2 with respect to the spatial context, two examples are shown in Figure 4 (a). People in the top F-formation standing side-by-side tend to have a large σ_2, while people in the bottom F-formation standing face-to-face tend to have a small σ_2.

From Table 1, for $T = 2/3$, our detector (spatial-**F**) shows competitive performance to the state-of-art. This is because tuning a global value of σ can already produce a good approximation of the clean F-formation shape, particularly as the soft detection threshold already considers partially detected members of an F-formation to be sufficient, enabling a softening of the need for strongly circular formations.

However, when considering the harsher criterion $T = 1$, our detector (spatial-**F**) significantly out-perform the state-of-the-art, suggesting that with a cross-validated comparison. We can also see that spatial-**F** detector performs equally good with both criterion $T = 2/3, 1$, which shows the accuracy of our detector is very high.

6.3. Results of Detecting Associates of F-formations

Table 2 shows that our proposed associate detector (social-**A**) significantly outperforms the three baselines (**SA**, **RA** and **ADA**), which means there are indeed certain patterns of associate behaviour that differs from the behaviour of singletons. We can also see from the performance **ADA** that it is also difficult for people to agree on who associates are. It also shows that social-**A(p+o)**with only proximity and orientation features can almost achieve the performance of complete set of features. Interestingly, global-**A** shows features extracted with less accurate F-formation detector can get the similar performance with social-**A** where more accurate F-formation detector spatial-**F** is used. This can be explained as our feature represents prototype-like F-formation structures, which can tolerant certain errors on less perfect F-formation detections.

To understand more about associates, some examples of them are shown in Figure 5. The red dots indicate the members' positions in an F-formation, the small red lines indicate everyone's orientation, the yellow dots indicate the correctly detected associates, the blue dots are correctly detected singletons, and the green dots show associates that were missed by the detector. From left to right, the first two images show that our detector can successfully detect associates who are in the r-space (See Figure 1(a)) trying to join an F-formation but who not accepted by its members. The third and fourth images show our detector can detect associates who are still in the F-formation p-space but not fully involved in the group. This conforms with the analysis of our analysis of the orientation and proximity of associates in Section 4 Figure 4(b).

We simulated tracking drifts on the manual labels of position and body orientation to compare the robustness of our method spatial-**F** with global-**F** on noisy test data. Figure 5 (b) shows that our context-aware F-formation detector spatial-**F** in general performs better than the detector with global parameters global-**F**, however, our detector can tolerate less noise by looking at the decay rate because our

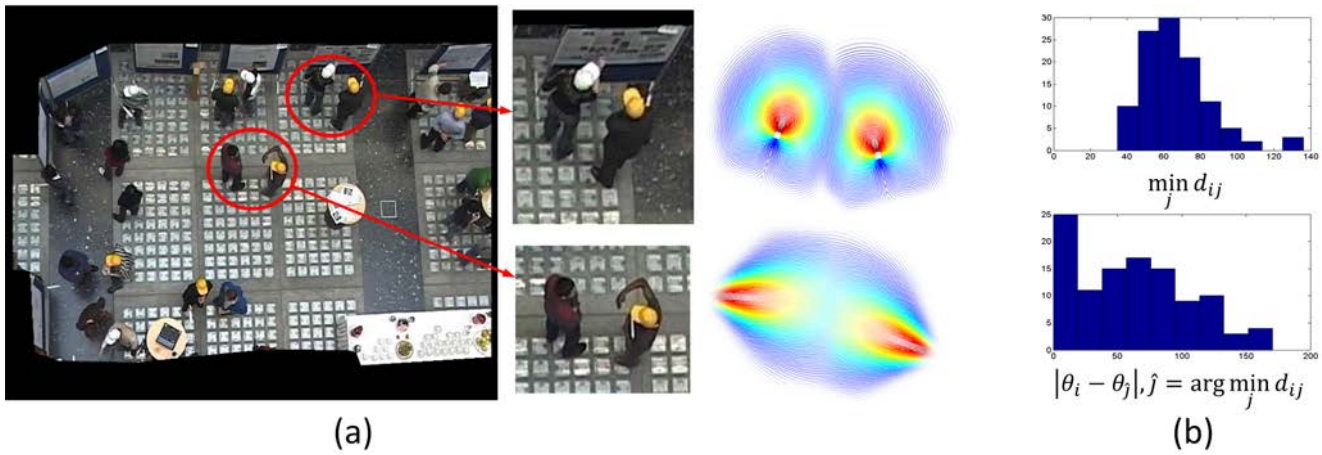

Figure 4. (a): learned frustum of attention in two cases. (b): histograms of both relative orientation differences between an associate and also distance to closest nearest F-formation member.

Figure 5. (a): example associate detection results: Red dots - members of an F-formation; red lines - body orientation; yellow dots - correctly detected associates; blue dots - correctly detected singletons; and green dots - missed associate detections. (b): F1 score of F-formation detectors spatial-**F** and global-**F** and associates detectors social-**A** and global-**A** with noisy test data.

learned parameters are sensitive to the location changing. As a person width is approximately 20 pixels in the image, the performance of our method starts to drop faster when the deviation of Gaussian noise is around half person width. It means our method performs good for reasonable robustness of trackers.

From Table 3, we can see that with feedback of detected associates, false positive F-formation members are removed, so that the precisions are improved significantly.

7. Conclusion

In this paper, we addressed the novel task of detecting associates of F-formations. We introduced a novel full multi-annotator set of annotations for associates of F-formations, and two methods for detecting them. Using our model, we were also able to discover patterns in proximity and orientation in the behaviours of associates that enable significant improvement over baseline methods with a detection rate of 71% F-measure. In terms of F-formation detection, We proposed a spatial-context-aware F-formation detector, which models people's frustum of attention in a principled way while considering the influence of the social and spatial context. The method is in general more adaptive to different datasets so for example, different frustum of attention parameters can be learned from scenarios with a non-uniform density of crowding. Our proposed method showed competitive performance , even when training the model parameters on less data.

By cleaning the detected in-group associates before re-performing F-formation detection, we were able to significantly improve F-formation detection on all cases where there was full-agreement amongst annotators on full-members of each F-formation. Surprisingly, althougth learning a spatial-context specific frustrum of attention led to better F-formation detection, when using the output of this models to detect associates, the performance for associate detection was not better than when F-formations were

detected with a spatial-context free frustrum parameters.

In summary, to our knowledge, this constitutes the first attempt on the challenging problem of automatically estimating conversational involvement levels in visual scenes of mingling.

References

[1] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(Mar):551–585, 2006.

[2] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino. Social interaction discovery by statistical analysis of F-formations. In *BMVC*, pages 1–12, 2011.

[3] M. Cristani, R. Raghavendra, A. D. Bue, and V. Murino. Human behavior analysis in video surveillance: A social signal processing perspective. *Neurocomputing*, 100(0):86–97, 2013. Special issue: Behaviours in video.

[4] T. Gan, Y. Wong, D. Zhang, and M. S. Kankanhalli. Temporal encoded f-formation system for social interaction detection. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 937–946. ACM, 2013.

[5] G. Groh, A. Lehmann, J. Reimers, M. R. Frieß, and L. Schwarz. Detecting social situations from interaction geometry. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 1–8. IEEE, 2010.

[6] H. Hung and B. Kröse. Detecting f-formations as dominant sets. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 231–238. ACM, 2011.

[7] N. Jovanović et al. Towards automatic addressee identification in multi-party dialogues. Association for Computational Linguistics, 2004.

[8] A. Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990.

[9] M. E. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[10] M. Pavan and M. Pelillo. Dominant sets and pairwise clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):167–172, 2007.

[11] L. Porzi, S. Rota Bulò, B. Lepri, and E. Ricci. Predicting and understanding urban perception with convolutional neural networks. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, MM '15, pages 139–148, New York, NY, USA, 2015. ACM.

[12] D. Santani and D. Gatica-Perez. Loud and trendy: Crowdsourcing impressions of social ambiance in popular indoor urban places. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, MM '15, pages 211–220, New York, NY, USA, 2015. ACM.

[13] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani. Multi-scale f-formation discovery for group detection. In *ICIP*, pages 3547–3551, 2013.

[14] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 22(12):1349–1380, 2000.

[15] 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*, 30(7):1212–1229, 2008.

[16] R. Subramanian, J. Staiano, K. Kalimeri, N. Sebe, and F. Pianesi. Putting the pieces together: multimodal analysis of social attention in meetings. In *Proceedings of the international conference on Multimedia*, pages 659–662. ACM, 2010.

[17] K. Tran, A. Gala, I. Kakadiaris, and S. Shah. Activity analysis in crowded environments using social cues for group discovery and human interaction modeling. *Pattern Recognition Letters*, 44:49–57, 2014.

[18] R. Triebel, K. Arras, R. Alami, L. Beyer, S. Breuers, R. Chatila, M. Chetouani, D. Cremers, V. Evers, M. Fiore, H. Hung, O. Islas Ramirez, M. Joosse, H. Khambhaita, T. Kucner, B. Leibe, A. Lilienthal, T. Linder, M. Lohse, M. Magnusson, B. Okal, L. Palmieri, U. Rafi, M. van Rooij, and L. Zhang. Spencer: A socially aware service robot for passenger guidance and help in busy airports. In *Conference on Field and Service Robotics (FSR)*, 2015.

[19] S. Vascon, E. Z. Mequanint, M. Cristani, H. Hung, M. Pelillo, and V. Murino. A Game-Theoretic Probabilistic Approach for Detecting Conversational Groups. In *ACCV*, 2014.

[20] S. Vascon, E. Z. Mequanint, M. Cristani, H. Hung, M. Pelillo, and V. Murino. Detecting conversational groups in images and sequences: A robust game-theoretic approach . *Computer Vision and Image Understanding*, pages –, 2015.

[21] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder. Bridging the gap between social animal and unsocial machine: A

survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87, 2012.

[22] N. Yasuda, K. Kakusho, T. Okadome, T. Funatomi, and M. Iiyama. Recognizing conversation groups in an open space by estimating placement of lower bodies. In *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*, pages 544–550, Oct 2014.

[23] T. Yu, S. Lim, K. A. Patwardhan, and N. Krahnstoever. Monitoring, Recognizing and Discovering Social Networks. In *CVPR*, 2009.