On Social Involvement in Mingling Scenarios
Detecting Associates of F-formations in Still Images

Zhang, Lu; Hung, Hayley

DOI
10.1109/TAFFC.2018.2855750

Publication date
2020

Document Version
Final published version

Published in
IEEE Transactions on Affective Computing

Citation (APA)
Zhang, L., & Hung, H. (2020). On Social Involvement in Mingling Scenarios: Detecting Associates of F-formations in Still Images. *IEEE Transactions on Affective Computing*, *12*(1), 165-176. Article 8413103. https://doi.org/10.1109/TAFFC.2018.2855750

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

On Social Involvement in Mingling Scenarios: Detecting Associates of F-Formations in Still Images

Lu Zhang and Hayley Hung

Abstract—In this paper, we carry out an extensive study of social involvement in free standing conversing groups (the so-called F-formations) from static images. By introducing a novel feature representation, we show that the standard features which have been used to represent full membership in an F-formation cannot be applied to the detection of so-called associates of F-formations due to their sparser nature. We also 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. Moroever, the majority of prior works have considered the labelling of conversing groups 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 can use the detected associates to improve initial estimates of the full members of an F-formation. We carry out extensive experimental validation of our proposed approach by collecting a novel set of multi-annotator labels of involvement on two publicly available datasets; The Idiap Poster Data and SALSA data set. Moreover, we show that parameters learned from the Idiap Poster Data can be transferred to the SALSA data, showing the power of our proposed representation in generalising over new unseen data from a different environment.

Index Terms—F-formations detection, human behaviour analysis, social group detection

1 Introduction

In recent years, the analysis of mingling scenarios has received growing attention [1], [2], [3], [4], [5]. With the recent advances in social signal processing [6], many potential applications of artificially intelligent perceptive systems are within reach. For example, potential applications include enabling robots to approach a group and offer assistance in a socially intelligent manner [7], or social surveillance [8], image interpretation or retreival [9].

Visual scene interpretation addresses the problem of bridging the semantic gap [9], 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 [10] or ambiance [11]. Similarly, in the area of social surveillance [8], researchers have been trying to ascribe social meaning to social scenes.

- *L. Zhang is with the Delft University of Technology, Delft 2628, CD, The Netherlands, and also with the University of Twente, Enschede 7522, NB, The Netherlands. E-mail: l.zhang@tudelft.nl.*
- *H. Hung is with the Delft University of Technology, Delft 2628, CD, The Netherlands. E-mail: h.hung@tudelft.nl.*

Manuscript received 26 Oct. 2017; revised 25 May 2018; accepted 1 July 2018. Date of publication 19 July 2018; date of current version 1 Mar. 2021. (Corresponding author: Hayley Hung.) Recommended for acceptance by F. Schwenker. Digital Object Identifier no. 10.1109/TAFFC.2018.2855750

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 and sometimes difficult to define, to bridge the gap in an informed manner, we exploit findings from social psychology to help inform how visually observed behaviours could be linked to social phenomena. Moreover, the inherent subjectivity in the perceptions of observed social behaviour provides an complexity to scene understanding. It brings new research challenges in understanding how to learn from and generate realistic models in the presence of differing but equally valid interpretations of the same visual data.

Given the great advances already in person tracking and orientation detection, in this paper, we focus on how to use their output as behavioural input for bridging the *social semantic gap*. Specifically, we approach the problem of detecting *associates* of conversing groups (or the so-called F-formation). F-formations are defined in psychology theory as a spatial organization of people gathered for conversation where each member has an equal ability to sense all other members [12]. 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 Fig. 1a). A more detailed definition is provided in the following section.

The majority of state-of-the-art methods for F-formation detection [5], [13], [14], [15], [16] have tended to make three simplifying assumptions. First, each individual is assumed to have a binary membership to an F-formation and to our

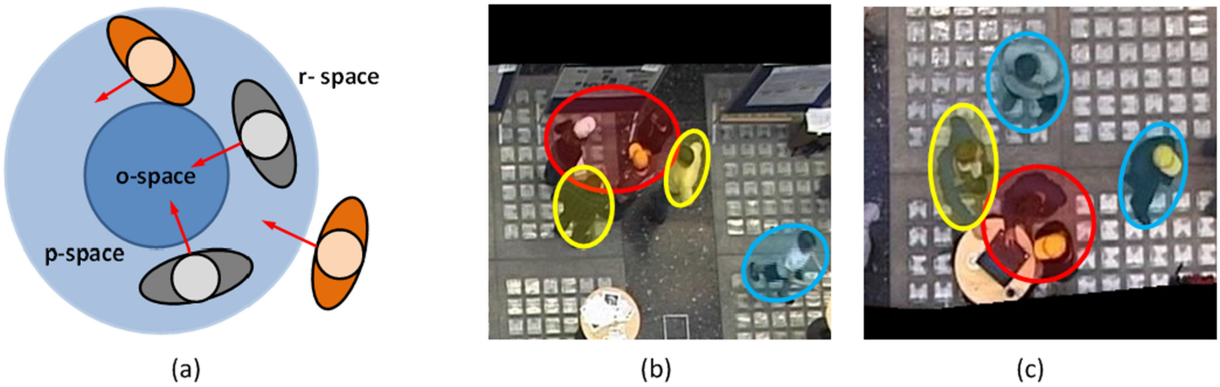

Fig. 1. Illustrations of F-formations. (a) The F-formation spaces, gray people stand in the p-space. Red arrows indicate body orientation. Orange people are associates of the F-formation. (b) and (c) Example snapshots: F-formations members, associates, and singletons are circled in red, yellow, and blue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

knowledge, no work has considered refining and enriching this model to label individuals who are partially involved in it. Bazzani et al. [17] proposed a solution that accounted for differing levels of group membership using a hierarchical tracking method. However, this was based purely on the distance between participants and could not distinguish involvement levels at a more subtle level.

In this paper, we show that the physical position of people can confuse the detection of full members of an F-formation as associates can stand in positions that are often mistaken for the behaviour of full members. 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 [12]. Finally, aside from Hung et al. [5], 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 and using these detections to correct for errors in the initial F-formation detection, we can also significantly improve performance on the F-formation detection task.

This paper makes 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 [15] 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 [5] for modeling associates. Finally, we carry out a deep evaluation and analysis of associates to investigate the complexity of this task.

An earlier version of this paper appeared in [18]. Compared to that work, this study contains (1) a substantial number of additional explanations and analysis, (2) various additional experiments to analyze and understand the results of the proposed method, and (3) a new multi-annotator labelling of the SALSA data, which we further validate our method on.

The outline of the remainder of this paper is as follows. We provide definitions of our study in Section 2, discuss related work in Section 3, and introduce the datasets used in this paper in Section 4. Section 5 described our new method of social involvement analysis on conversational groups. We present the results of our experiments in Section 6. Section 7 concludes the paper.

2 Definitions

2.1 F-Formations and Their Associates

The psychologist Kendon [12] 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 Fig. 1a. 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 [5], [15], [16]. 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 [12]. For instance, people talking in front of a laptop may stand closer and look at the same direction (see Fig. 1c), which maintains an F-formation although their o-space could be violated.

Unlike full members of F-formations, Kendon [12] 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 [12].

While F-formations can easily be modelled by either maximal cliques [5], [15], [16] or a joint centre-of-focus in the o-space [13], associate behaviours are not so clearly linked to a single set of social cues. Therefore, the associate detection problem requires us to bridge 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 other members of the groups. Meanwhile, associates have the least potential to influence full members but could be influenced by them. Crucially, in-group associates could be mistaken for full F-formation members and out-group associates for singletons.

2.2 Frustum of Attention

The frustum of attention [15] (or transactional segment, as defined by Kendon [12]) 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], [19], [20], body pose [5], gaze [20], [21], and proximity [5] 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 [13], [14], [15]. There are two main drawbacks of this approach. First, the parameters are likely to over-fit 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 [15], [19], [20], body pose [1], gaze [20], [21], and proximity [5] 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 [23] was proposed to identify F-formations from automatically extracted trajectories by [24]. To our knowledge, in terms of the treatment of hierarchy in groups, the work of [24] 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 [13], which estimates the locations of o-spaces by density estimation. The size of F-formation was taken into account using a multiscale Hough voting strategy in [14]. In [5], [15], 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 [25] in the graph and the edges of the graph are computed based on body orientation and proximity. In [15], the temporal information is added in the dominant set based approach. A density-based approach was proposed in [26] where the final purpose of the task was to dynamically select camera angles for automated event recording. In [27], temporal patterns of activities were subsequently analyzed. In this paper, we follow the dominant set framework because it gives reliably good results in general [15] 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

4.1 Idiap Poster Data

We used the publicly available Idiap Poster Data [5],[1] which consists of 3 hours of aerial video of over 50 people 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 Fig. 7. In total, 82 images including 1,700 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. [5]. We augmented this data by adding annotations of associates of the F-formations.

We analyzed the annotations[2] to see whether there was full agreement between the annotators about all members of an F-formation and associates. 211 instances of associates were annotated. 84 associates were identified with majority agreement (39.8 percent) and 34 for full agreement (16 percent). 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 percent respectively, which shows that associates are not as straight forward to label compared to F-formations (94.74 percent mean average F-measure when computing the agreement for F-formations from the data). We consider all the annotated associates can have different levels of involvement with respect to their associated F-formation. To have a intuitive feeling of agreement among 3 annotators, we also compute the traditional kappa statistic. The average Kappa of

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

2. The annotations generated by this paper are available to research institutions subject to an End User License Agreement by contacting h.hung@tudelft.nl.

3 pairs of annotators (AB,AC, and BC) is 0.42. The computation is simply based on checking if two annotators agree on an individual sample to be an associate or not. We can see that the agreement on average is not high, which also shows the complexity of annotating the associates together with F-formation due to the subjectiveness.

To explore the relative angle and orientation relationship between different types of associates of F-formations, we computed histograms of both the distance to, and the relative orientation differences between, an associate and his closest F-formation member as shown in the top and bottom of Fig. 7b on p. 22 respectively. The relative orientation of associates to their closest F-formation member has a peak in probability mass at 0, and $\pi/3$ while there is only a single peak in the lower histogram. This shows that associates tend to stand similarly closely to their nearest F-formation member.

The double peak seen in the relative orientation suggests that the idea of two types of associates may be true. Those who stand in the p-space of an F-formation but appear less involved in the conversation (in-group associates) could be representative of the peak at the first bin where there is almost no difference in orientation while those that stand in the r-space, facing towards the F-formation (out-group associates) could be formed from the remainder of the samples populating the histogram. To show this definitively, we would need to have a accurate model of the o-, p-, and r-space. Unfortunately, the spatial layout of the room means that the circular formation will be distorted. It is left for further research to define robust models of these spaces with respect to proximity, orientation, and differences in annotator agreement.

4.2 SALSA Dataset

We also used the SALSA dataset [2][3] to test the robustness of our method, which consists of two parts (poster session and coffee break). Each part lasts for approximately half an hour was recorded by four side view cameras pointed towards the area of interest as shown in Fig. 3. In this paper we only use the data during the poster session since people are mingling more between different groups whilst in the coffee break most people surround a table just waiting for drinks. This enables more different group formations and also examples of associates to appear and be evaluated on. 18 people participated during the poster session. In the room was 4 poster boards arranged around the perimeter of the scene and one table where refreshments could be taken. The dataset has an existing set of annotations of people's location on the ground plane, body/head orientation, and F-formations based on automated detections.

We re-annotated the data for F-formation and associates with 3 annotators because 1) the annotation provided by the dataset contains only 1 annotation for each F-formation, 2) there are no annotations of associates, and 3) the F-formation annotations provided by the data set were labelled based on automated detections and not real locations.

Similar to the Idiap Poster Data, the positions of all the people in the scene were pre-labelled so that the annotators could concentrate on identifying the F- formations. Software was written to allow easy labelling of the data. Three annotators from different international cultures (American, Chinese, and Dutch) and professional backgrounds volunteered to label the data and were remunerated for their efforts with a gift voucher. The annotators were asked to label the same data so that variability in the labelling could be taken into account during evaluation. The annotators were asked to label F-formations and their associates after an initial training phase where definitions were given to them for each type of person. Annotators were given an initial training phase with appropriate definitions for F-formations and their associates before they started labelling the images. Asking for explicit labels for associates ensured that annotators would consciously decide how involved they thought each person was in the corresponding F-formation. A snapshot is shown in Fig. 4, where the pre-labelled positions of every persons head are indicated by a yellow box. On screen instructions help the annotators to singletons, identify the F-formations and label its corresponding associates, which are subsequently highlighted with a corresponding colour code.

The three annotators given pre-selected image frames during the poster session approximately every 2-3 minutes, resulting in 13 images and 234 individual instances of people to be labelled. Using our new annotations, and like the Idiap Poster Data, we studied the annotator agreement. We analyzed the annotations to see whether there was full agreement between the annotators about all members of an F-formation and any corresponding associates. The number of full agreement F-formations was 192, and 214 when the union of labelled full members was used. 37 instances of associates were annotated, where 20 were identified with majority agreement (54.1 percent) and 14 with full agreement (37.8 percent). We computed the F1 score considering one annotation as ground truth and another annotation as a detection for each set of data annotated by the same 3 annotators. The mean and standard deviation of the F1 score was 67.4 and 5.6 percent respectively. We can attribute the higher annotator agreement to the more constrained scenario—with just 18 people, and the fact that we studied the poster session and not a mixture of poster and mingling behaviour like for the Idiap Poster Data, it is logical that we would see a higher annotator agreement on the associates in this case.

5 Methodology

We detect an associate by modeling its social prior with its associated conversational group (F-formation) based on non-verbal cues 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 Fig. 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; trained classifiers will be used to determine the involvement of a certain person to a F-formation, for instance, F-formation members, associates, or singletons. The modules are described in the following sections separately.

5.1 Modeling the F-Formation as a Dominant Set

Building on prior work [5], [15], we exploit the dominant set framework. In an image, people can be represented as a

3. http://tev.fbk.eu/salsa

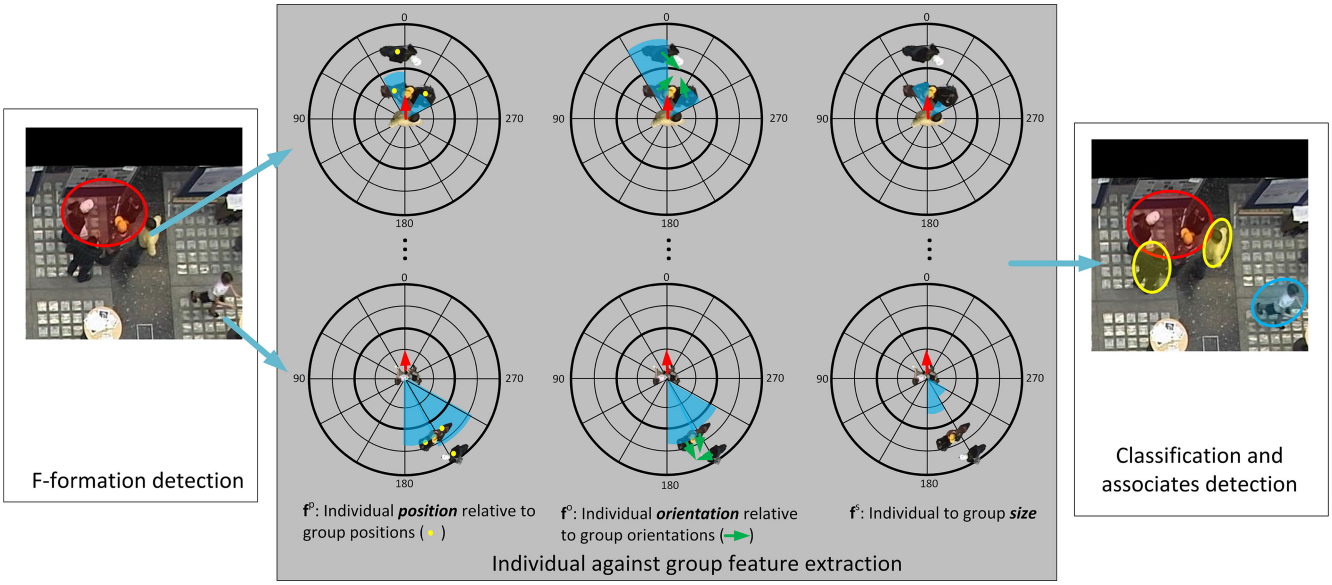

Fig. 2. Flow diagram showing the stages of F-formation and associate detection.

Fig. 3. Screenshots of four cameras of SALSA dataset.

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 an affinity function which defines the "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 is defined as

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

which measures the overall relative affinity between i and the rest of the nodes in S and where $R = S \backslash \{i\}$. As described in [25], the relationship between internal and external nodes of a dominant set S are conditioned on

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

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

Detecting a dominant set is identical to solving the following standard quadratic programme

$$\max_{\mathbf{x}} \mathbf{x}^T A \mathbf{x}, \quad s.t. \ \mathbf{x} \in \Delta, \tag{4}$$

where the standard simplex

$$\Delta = \left\{ \mathbf{x} \in R^{|V|} : \sum_{i \in V} x_i = 1, \ x_i \geq 0, \ i = 1, \ldots, |V| \right\}. \tag{5}$$

This optimization problem can be solved with a method from evolutionary game theory, called replicator dynamics. The first-order replicator can be represented as

$$x_i = x_i \frac{(A\mathbf{x})_i}{\mathbf{x}^T A \mathbf{x}}. \tag{6}$$

Once $\mathbf{x}$ converges, one set of F-formation members are detected. Assuming we have $|V|$ independent strategies, the pay-off of strategy i play against strategy j is a_{ij}. In time t, the probability of picking up strategy i is x_i. The average pay-off of strategy i at time t is $(A\mathbf{x}(t))_i$ and the average pay-off of all the strategies at time t is $\mathbf{x}(t)^T A \mathbf{x}(t)$. The main idea is that over time, bad strategies will die off and stable strategies will last. If we consider people in our application as strategies, the selected people in the end are the optimal solution of a dominant set. That is, after the game converges, the indexes of non-zero x_i identify the members of the F-formation.

This strategy only identifies one F-formation at a time. To identify all of them in the graph, a peel strategy is used. This means that once all non-zero elements of $\mathbf{x}$ are identified, nodes associated with these elements are removed from the graph and the optimization procedure is repeated in the remainder of the graph. 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. Similar to [5] this enables a stopping criterion that is sensitive to the global context of the scene where the number of clusters to

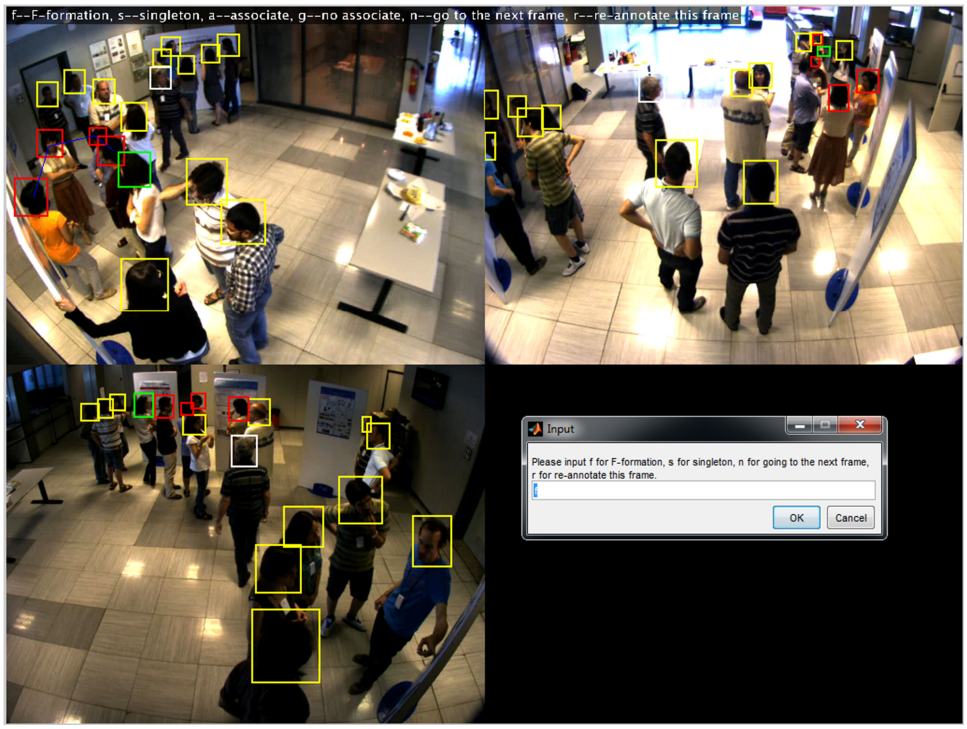

Fig. 4. Example snapshot of the annotation graphical user interface where one F-formation has been annotated for. Labelled F-formation full members, associates, and singletons are indicated by red, green and white boxes respectively. The blue lines (top right frame) between the red boxes indicate which F-formation the full members belong to.

be found does not need to be determined beforehand. For more details, see [5], [25].

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 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 an associate and F-formation is NP. Fortunately, in practice, associates tend to be scatted sparsely enough amongst the F-formations in a scene so that the maximal clique assumption for a single F-formation 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 incorrect F-formation detection in the earlier step (e.g., detecting associates as full F-formation members), every individual in the data is considered as an associate candidate, so an associate candidate could be an F-formation member, an associate, or a singleton in reality. 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 the kth F-formation member in C.

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 an associate candidate $\angle(\mathbf{p}_k - \mathbf{p}_a)$, so that every bin covers an angle of $\pi/6$. We define the mth bin of the three sets of features as

$$\mathbf{f}^p_m = \frac{1}{Z_d \cdot |C_m|} \sum_{k \in C_m} \left\| \mathbf{p}_k - \mathbf{p}_a \right\|, \tag{7}$$

$$\mathbf{f}^o_m = \frac{1}{Z_o \cdot |C_m|} \sum_{k \in C_m} \left(\angle \mathbf{p}_k - \angle \mathbf{p}_a \right), \tag{8}$$

$$\mathbf{f}^s_m = \frac{1}{Z_s} |C_m|, \tag{9}$$

where the set of F-formation members located in this bin is C_m. We use $\mathbf{f}^p_m$ to represent the average distance between F-formation members in C_m and $\mathbf{p}_a$, $\mathbf{f}^o_m$ to represent the average relative body orientation between F-formation members in C_m and $\mathbf{p}_a$, and $\mathbf{f}^s_m$ to represent the relative person 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 Fig. 2

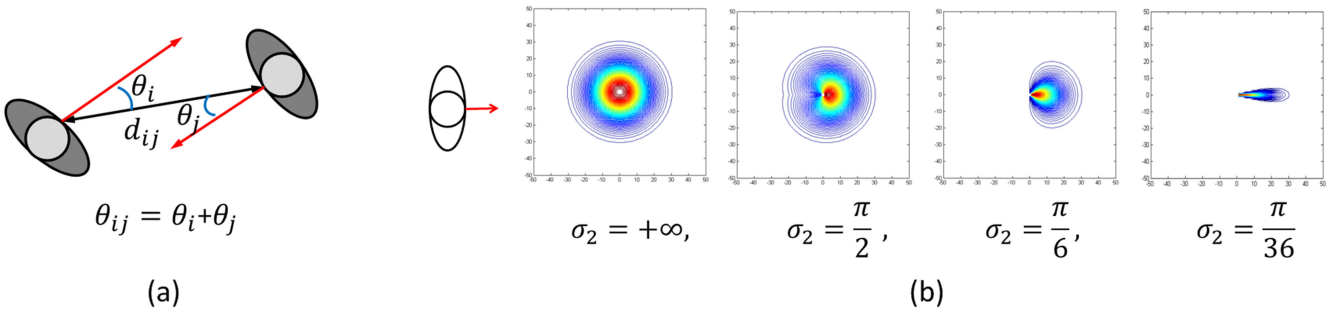

Fig. 5. Frustum of attention modeling with body orientation and proximity. (a) Calculation of relative orientation and proximity, (b) frustum of attention map with different parameters. The smaller the σ_2 is, the narrower of frustum attention of a person is.

shows examples of the scale or orientation invariant feature representations of an associate and a singleton, which encode people's relative location, orientation and group size.

Associatess detection is challenging because they are likely to be detected as full F-formation members compare to singletons who are usually far away from an F-formation. We use a one-versus-the rest strategy to train an associates detector. In our experiment, we used 211 instances of annotated associates, 235 full-agreement singletons and 450 full-agreement F-formations as training data. In the experiment, we compare a set of classifiers; Parzen, RBF SVM, Random Forests, and AdaBoost, with 10 fold cross validation. The Parzen classifier gave the best performance on our dataset, therefore, we have chosen to use this classifier for the remainder of the experiments in this paper.

5.3 Training the 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

$$a_{ij} = e^{-\frac{d_{ij}^2}{\sigma_1^2} - \frac{\theta_{ij}^2}{\sigma_2^2}}, \tag{10}$$

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 Fig. 5), 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 Fig. 7a). 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 Fig. 7a). The objective function is defined as

$$\ell = \sum_{n=1}^{N} 1 - \frac{|C^{\{n\}} \cap \hat{C}^{\{n\}}|}{|C^{\{n\}} \cup \hat{C}^{\{n\}}|}, \tag{11}$$

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. During training, we consider a detection C and an annotation $\hat{C}$ to match with each other if $\frac{|C \cap \hat{C}|}{|C \cup \hat{C}|} \geq \frac{2}{3}$. 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 only update the parameters once per person when the detection goes wrong in a passive-aggressive way [28]

$$\sigma_s(\mathbf{p}) = \sigma_s(\mathbf{p}) - g_s(C)\Delta\sigma_s, \quad s \in \{1, 2\}. \tag{12}$$

Here, $\Delta\sigma_s$ is the basic step size, which is set to a small value ($\Delta\sigma_s = 0.1$ in our experiment). An adaptive parameter g helps to adapt to different F-formation geometric variations. Given F-formation C, the adaptive parameter g is defined as

$$g_1(C) = y\frac{\|\sum_{i,j\in\hat{C}^{\{n\}}} \hat{d}_{ij} - \sum_{i,j\in C^{\{n\}}} d_{ij}\|}{\sum_{i,j\in\hat{C}^{\{n\}}} \hat{d}_{ij}}, \tag{13}$$

$$g_2(C) = y\frac{\|\sum_{i,j\in\hat{C}^{\{n\}}} \hat{\theta}_{ij} - \sum_{i,j\in C^{\{n\}}} \theta_{ij}\|}{\sum_{i,j\in\hat{C}^{\{n\}}} \hat{\theta}_{ij}}, \tag{14}$$

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

6 Experimental Results

We performed three sets of experiments to evaluate the performance of our system; F-formation detection, associates detection, and improved F-formation detection using the feedback of associates detection. We did the experiments on two datasets (Idiap Poster and SALSA), which to our, knowledge, are currently the largest publicly available datasets with multi-person annotations. In this section, we first describe the experiment on the Idiap Poster Data in Sections 6.1, 6.2, and 6.3, then we study the feature representation in Section 6.4, finally, the study using the SALSA dataset where we trained parameters on the Idiap Poster Data are provided in Section 6.5.

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 number of iterations 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

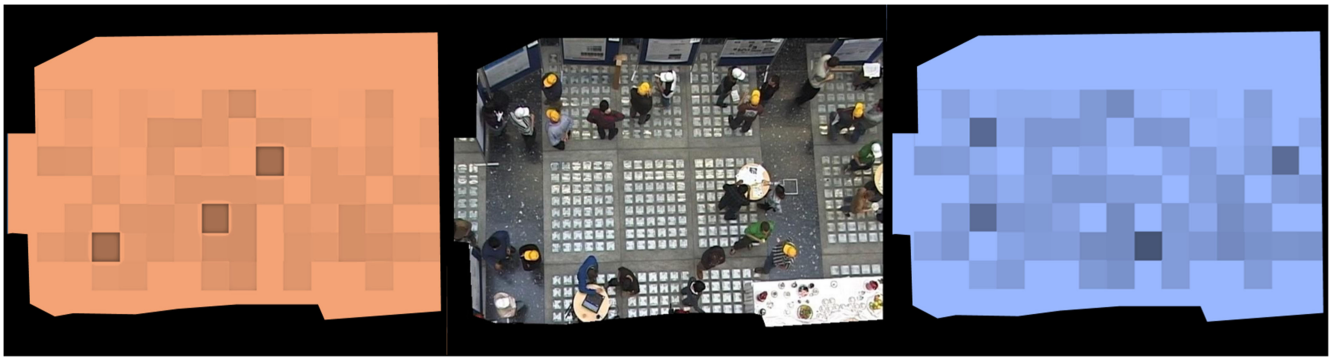

Fig. 6. Learned values of σ_1 (left) and σ_2 (right) spatially on the image plain.

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. [5].

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 [13], 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 the 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 spatially-aware F-formation detector (Spatial-**F**) with state-of-the-art *DSFF* [5], *HFF* [13], *ACCVKL* [15], and *ACCVJS* [15].

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: proximity features, o: orientation features, and s : group size features). The associates detector global-**A** extracts features based on global-**F** F-formation detection.

Finally, we analysed how associate detection can help improve F-formation detection. As the F-formation detector has problems mostly with in-group associates, we used the detected associates to clean up false positives in a detected F-formation. The performance of Spatial-**F** and global-**F** was evaluated with the $T = 1$ hard criterion using F-formations annotated with full agreement.

6.2 F-Formation Detection Results

Two examples of the learned values for σ_1 and σ_2 with respect to the spatial context, are shown in Fig. 7a. 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. An over all learned σ_1 and σ_2 over the image plain is shown in Fig. 6, where the range of learned values are $\sigma_1 \in [2802, 3600]$ and $\sigma_2 \in [\frac{2\pi}{5}, \frac{35\pi}{36}]$. We can see there are a few dark blocks from the learned σ_1 map, Which means that the distance between people plays a more important role for detecting F-formation in the empty area (a round and small circular F-formation is easily formed); similarly, the learned σ_2 map tells us that orientation is more important for detecting F-formations on the sides (in front of posters, a flat F-formation often appears). In addition, we can see that the maps are a bit sparse due to the sparsity of our training data in the space. Therefore, the parameter at a certain location might not be learned due to a lack of data in such an area.

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-performs the state-of-the-art, even with a cross-validated comparison. We can also see that the spatial-**F** detector performs equally good with both criteria ($T = 2/3$ or 1), which shows the accuracy of our detector is very high.

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 [5]	0.93	0.92	0.92	0.81	0.81	0.81
HFF [13]	0.93	0.96	0.94	0.81	0.84	0.83
ACCVKL [15]	0.90	0.94	0.92	-	-	-
ACCVJS [15]	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

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

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 Section 6.1).*

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 when using all the features (social-*A (p+o+s)*). Interestingly, global-**A** shows features extracted with a less accurate F-formation detector can still obtain a similar performance with social-**A** where a more accurate F-formation detector spatial-**F** was used. This can be explained as our feature represents prototype-like F-formation structures, which can tolerate certain errors on less perfect F-formation detections.

To understand more about associates, some examples of them are shown in Fig. 8. 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 Fig. 1a) trying to join an F-formation but who are not accepted by its members. The third and fourth images show that our detector can detect associates who are still in the F-formation p-space but not fully involved in the group. This conforms our analysis of the orientation and proximity of associates in Section 4 Fig. 7b.

As described in Section 4, we now explore the relative angle and orientation relationship between different types of associates of F-formations. To do this, we computed the histograms of both relative orientation differences between an associate and also distance to closest their nearest F-formation

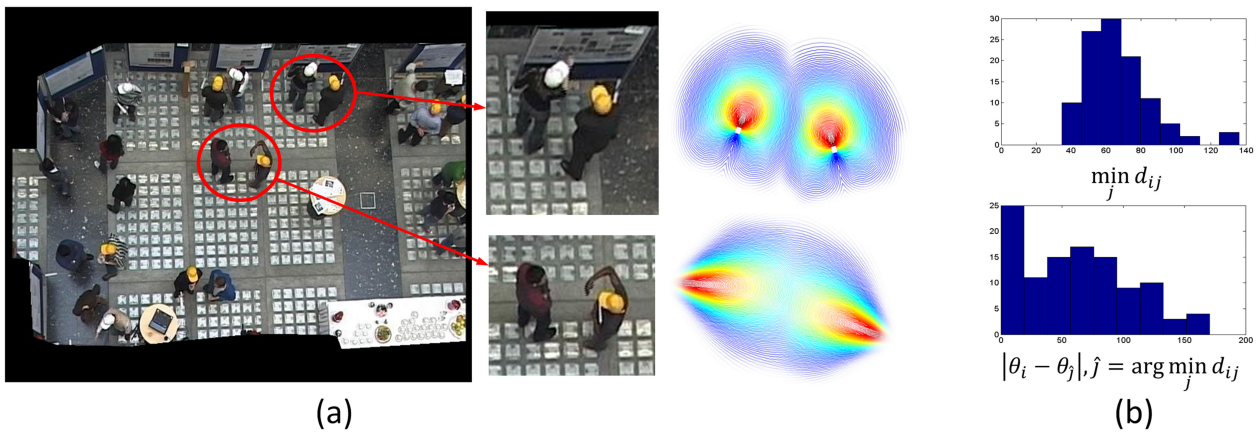

Fig. 7. (a): Learned frustum of attention in two cases. (b): Histograms of both relative orientation differences between an associate and also distance to closest nearest F-formation member.

Fig. 8. (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.

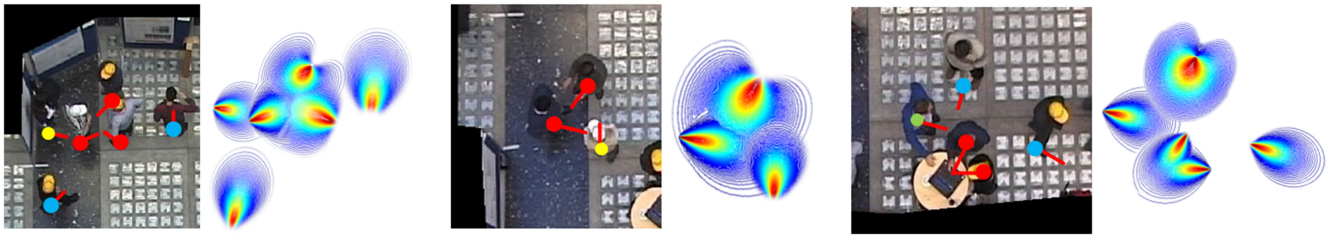

Fig. 9. Learned frustra of attention of three scenes. See caption of Fig. 8 for legend of dots and lines in images. Associated frustra of attention are shown on right of each image.

member as shown in the top and bottom parts of Fig. 7b 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). Based on our detection results, we use a threshold $\pi/6$ to see roughly how good our detector is for detecting in/out group associates. We calculated the number of correct detected associates, the in/out group ratio is 1:1.44, compared to ground truth with the same threshold 1:1.41, the results shows that our social involvement features can represent both in and out group associates very well.

To further explain the success and failure cases with respect to the trained parameter values, the learned frustum of attention maps of two associates and one failure case are shown in Fig. 9. We can see that people in F-formations stand in the higher attention area of other full members; associates tend to cover F-formation members in their higher attention area but are not themselves in the higher attention area of all members of the F-formation; singletons are not in anyone's higher attention area, and do not tend to have people standing in their higher attention areas.

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. Fig. 8b shows that our 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 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 should perform well using a reasonably robust visual tracker.

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

TABLE 3
F-Formation Detection with Associate Detection Feedback, Results Are Evaluated Only on F-Formations Annotated with Full-Agreement

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

FB-global-F and FB-spatial-F are detectors with associate detection feedback (details in Section 6.1).

6.4 Full and Associate Member Feature Analysis

To highlight the novelty of our proposed associate representation further, we also provide analysis to show how we match the proposed feature representations to the two problems: F-formation and associates detection. Full members of the F-formations are represented by a single expression (Eq. (10)). This allows us to jointly learn the angle and length of the view frustum. Rather than using the same formulation to represent associates, we identified that a different rotation, position, and size invariant representation of associates was needed (Eqs. (7), (8), and (9)) due to their sparse nature. This duality is a subtle but important point which is best demonstrated in the F1 scores in the table below where we show how the performance changes if we use the features designed for associates on full member detection or vice versa. We see clearly that full-members are better learned from the dense representation of Eq. (10) while associates are better learned from the sparse representations of Eqs. (7), (8), and (9). Our results show that matching the right representation to the right problem performs much better demonstrating further the very different nature of the associate problem compared to the traditional F-formation full-member detection problem.

6.5 Demonstrating a Generalised Model of Associates on the SALSA Dataset

As described in Section 4, the number of both annotated F-formations and associates in the SALSA dataset are much less than those annotated in the Idiap Poster Data (37 annotated associates in the SALSA dataset versus 211 in the Idiap Poster Data). To avoid overfitting on such a small dataset, we used the trained associate detectors from the Idiap Poster Data to test the generalisation on the unseen SALSA

TABLE 4
Feature Representation Comparison with F1 Measure

	Eq. (10): tight representation	Eqs. (7), (8), and (9): sparse representation
F-formation detection (tight problem)	0.94	0.65
Associate detection (sparse problem)	0.31	0.71

TABLE 5
F-Formation and Associates Detection Result on SALSA Data

Method	Prec.	Rec.	F1
global-F $[T = 2/3]$	0.82	0.93	0.87
global-F $[T = 1]$	0.56	0.64	0.60
FB-global-F $[T = 2/3]$	0.83	0.93	0.87
associate detection	0.63	0.67	0.64

data. Since the room arrangements are different for these two datasets, we did not use the location prior trained using the Idiap Poster Data, but a set of fixed parameters $\sigma_1 = 40, \sigma_2 = 30$ across the entire scene. The F-measure of the F-formation detection, associates detection, and F-formation detection with feedback from associates detection on this data are shown in Table 5.

Note that since the F-formations and corresponding associates were re-annotated with multiple new annotators, it is not meaningful to compare the performance of our method with that reported by Alameda-Pineda et al. [2] since their ground truth is essentially different. However, given the higher annotator agreement and less complex set-up compared to the Idiap Poster Data, we would expect to see comparable F-formation detection results as we observed with the Idiap Poster Data. As with the earlier setup of the experiments on the Idiap Poster Data, we evaluated the F-formation detections with full-agreement annotations, and the union of all labelled associates.

The results shows that with global-F detector, we can still achieve reasonable F-formation detection accuracy with $T = 2/3$, however, with a more strict criterion $T = 1$, detecting without learning the spacial prior shows much less accuracy given most of the F-formations have a side-by-side pattern. As a result, in this case, we see no further improvement of the full-member detection when exploiting the associate detection as feedback. Applying our learned associates model from the Idiap-poster data to the SALSA data, we still captured the associates with an F-measure of 0.64. This demonstrates that the we have learned a relatively general model of associates that can be learned from one scene and applied to another.

7 Conclusion and Discussion

In this paper, we addressed the task of automatically detecting social involvement in conversing groups. Specifically, we studied the detection of associates of F-formations, validating on two publicly available data sets of natural human behaviour. We introduced a novel full multi-annotator set of annotations for associates of F-formations for the publicly available SALSA data set, 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 percent 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.

Experiments on the more complex Idiap Poster Data showed that by cleaning the detected in-group associates before re-performing F-formation detection, we were able to significantly improve F-formation detection in 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. Moreover, we show that a significantly different representation of associates must be learned compared to those of full members of F-formations.

Finally, our experiments with the SALSA data set show that our method is able to learn a generaliseable representation of associates that can be applied on a completely different data set.

In terms of future work, while this paper provided the ground work to understand the nature of the associate detection problem better, the next step would be to use automated detections. This leads to quite a number of issues caused by error propagation with respect to false detections, missed detections, or difficulties in disambiguating the pixels of one person from another. One approach [13] used to mitigate this problem is to only allow annotators to provide labels for people (heads) who are automatically detected. However, then there is clearly the possibility of bias due to people, possibly associates being missed completely during the detection phase and thus never annotated for. Some works have also performed automated analysis with some form of association of the detections to the ground truth. However, the precise rules of association are not stated or discussed [2]. One way to mitigate the problems is to process only frames in which all participants are visible [2]. However, this still does not help in handling situations of varying occlusion.

Another shortcoming is that this work is based on static observations. Many real life applications will have access to video. It is also highly likely that the perception of involvement might change as a result of observing the interaction dynamics of a conversation. Going back to the above discussion about automated detection of the position and orientation, over time, people must be tracked and therefore associated between frames. If there is heavy occlusion, deciding decisions need to be made about how to associate tracks together and to the same person in an F-formation over time is still an unaddressed challenge.

Finally, as discussed in [29], often with fully automated methods, the head pose is used as a proxy for body orientation since it tends to be less occluded. In an ideal case, to estimate the F-formation itself, the body orientation is more descriminative. On the other hand, the head pose should indicate the dynamics of the conversational structure, which could serve to indicate the conversational involvement better. Therefore, in moving forward with more dynamic analyses, one should be mindful of the role that both head and body can play.

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. We hope this spurs further work to investigate some of the issues discussed above so we can move smart surveillance systems to become more socially intelligent.

Acknowledgments

This work has partly been supported by the European Commission under contract number FP7-ICT-600877 (SPENCER). The authors thank Jan van Gemert and Julian Kooij for their helpful discussions during the preparation of this work.

References

[1] G. Groh, A. Lehmann, J. Reimers, M. R. Frieß, and L. Schwarz, "Detecting social situations from interaction geometry," in *Proc. IEEE 2nd Int. Conf. Social Comput.*, 2010, pp. 1–8.
[2] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe, "SALSA: A novel dataset for multimodal group behaviour analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1707–1720, Aug. 2016.
[3] F. Setti, C. Russell, C. Bassetti, and M. Cristani, "F-formation detection: Individuating free-standing conversational groups in images," *PloS One*, vol. 10, no. 5, 2015, Art. no. e0123783.
[4] E. Ricci, J. Varadarajan, R. Subramanian, S. R. Bulo, N. Ahuja, and O. Lanz, "Uncovering interactions and interactors: Joint estimation of head, body orientation and F-formations from surveillance videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4660–4668.
[5] H. Hung and B. Kröse, "Detecting F-formations as dominant sets," in *Proc. 13th Int. Conf. Multimodal Interfaces*, 2011, pp. 231–238.
[6] 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 Trans. Affective Comput.*, vol. 3, no. 1, pp. 69–87, Jan.–Mar. 2012.
[7] 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 *Proc. Conf. Field Service Robot.*, 2015, pp. 607–622.
[8] M. Cristani, R. Raghavendra, A. D. Bue, and V. Murino, "Human behavior analysis in video surveillance: A social signal processing perspective," *Neurocomput.*, vol. 100, pp. 86–97, 2013.
[9] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.
[10] L. Porzi, S. Rota Bulò, B. Lepri, and E. Ricci, "Predicting and understanding urban perception with convolutional neural networks," in *Proc. 23rd Annu. ACM Conf. Multimedia Conf.*, 2015, pp. 139–148. [Online]. Available: http://doi.acm.org/10.1145/2733373.2806273
[11] D. Santani and D. Gatica-Perez, "Loud and trendy: Crowdsourcing impressions of social ambiance in popular indoor urban places," in *Proc. 23rd Annu. ACM Conf. Multimedia Conf.*, 2015, pp. 211–220. [Online]. Available: http://doi.acm.org/10.1145/2733373.2806277
[12] A. Kendon, *Conducting Interaction: Patterns of Behavior in Focused Encounters*, vol. 7. Cambridge, U.K.: Cambridge Univ. Press, 1990.
[13] 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 *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 1–12.
[14] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani, "Multi-scale F-formation discovery for group detection," in *Proc. IEEE Int. Conf. Image Process.*, 2013, pp. 3547–3551.
[15] S. Vascon, E. Z. Mequanint, M. Cristani, H. Hung, M. Pelillo, and V. Murino, "A game-theoretic probabilistic approach for detecting conversational groups," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 658–675.
[16] 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," *Comput. Vis. Image Understanding*, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314215002076
[17] L. Bazzani, M. Zanotto, M. Cristani, and V. Murino, "Joint individual-group modeling for tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 746–759, Apr. 2015.
[18] L. Zhang and H. Hung, "Beyond F-formations: Determining social involvement in free standing conversing groups from static images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1086–1095.
[19] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1212–1229, Jul. 2008.
[20] R. Subramanian, J. Staiano, K. Kalimeri, N. Sebe, and F. Pianesi, "Putting the pieces together: Multimodal analysis of social attention in meetings," in *Proc. Int. Conf. Multimedia*, 2010, pp. 659–662.
[21] N. Jovanović, and H. J. A. op den Akker, "Towards automatic addressee identification in multi-party dialogues," in *Proc. 5th SIGdial Workshop on Discourse and Dialogue*, Boston, MA, pp. 89–92, 2004.
[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 *Proc. IEEE Int. Conf. Syst. Man Cybern.*, Oct. 2014, pp. 544–550.
[23] M. E. Newman, "Modularity and community structure in networks," *Proc. Nat. Academy Sci.*, vol. 103, no. 23, pp. 8577–8582, 2006.
[24] T. Yu, S. Lim, K. A. Patwardhan, and N. Krahnstoever, "Monitoring, recognizing and discovering social networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1462–1469.
[25] M. Pavan and M. Pelillo, "Dominant sets and pairwise clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 167–172, Jan. 2007.
[26] T. Gan, Y. Wong, D. Zhang, and M. S. Kankanhalli, "Temporal encoded F-formation system for social interaction detection," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 937–946.
[27] 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 Recognit. Lett.*, vol. 44, pp. 49–57, 2014.
[28] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *J. Mach. Learn. Res.*, vol. 7, pp. 551–585, 2006.
[29] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe, "Analyzing free-standing conversational groups: A multimodal approach," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 5–14.

Lu Zhang received the PhD degree in computer vision from the Delft University of Technology, The Netherlands, in 2015. Her PhD thesis was about model-free tracking as well as object, gesture, and action recognition. She is currently a postdoc with the Pattern Recognition & Bioinformatics Group, Delft University of Technology and University of Twente. Her research interests include social signal processing and computer vision.

Hayley Hung received the PhD degree in computer vision from the Queen Mary University of London, United Kingdom, in 2007 and her first degree from Imperial College, United Kingdom, in electrical and electronic engineering. She is an assistant professor and Delft Technology fellow with the Pattern Recognition and Bioinformatics Group, TU Delft, The Netherlands, since 2013. She was awarded the Dutch Research Foundation (NWO) Career Talent Award for experienced researchers (Vidi) in 2015. Between 2010-2013, she held a Marie Curie Intra-European Fellowship with the Intelligent Systems Lab, University of Amsterdam. Between 2007-2010, she was a post-doctoral researcher with Idiap Research Institute in Switzerland. Her research interests include social computing, social signal processing, computer vision, and machine learning.