Detecting F-formations as Dominant Sets

Hayley Hung
Intelligent Systems Lab
University of Amsterdam
H.Hung@uva.nl

Ben Kröse
University of Amsterdam and
Hogeschool van Amsterdam
b.j.a.krose@hva.nl

ABSTRACT

The first step towards analysing social interactive behaviour in crowded environments is to identify who is interacting with whom. This paper presents a new method for detecting focused encounters or F-formations in a crowded, real-life social environment. An F-formation is a specific instance of a group of people who are congregated together with the intent of conversing and exchanging information with each other. We propose a new method of estimating F-formations using a graph clustering algorithm by formulating the problem in terms of identifying dominant sets. A dominant set is a form of maximal clique which occurs in edge weighted graphs. As well as using the proximity between people, body orientation information is used; we propose a socially motivated estimate of focus orientation (SMEFO), which is calculated with location information only. Our experiments show significant improvements in performance over the existing modularity cut algorithm and indicates the effectiveness of using a local social context for detecting F-formations.

1. INTRODUCTION

Automatically estimating relationships between humans is a challenging problem, being a highly varied and subjective phenomenon which is difficult to categorise and capture. By relationships, we refer to behavioural phenomena such as dominance, interest, or attraction which can be perceived during prolonged conversation. Typically, such behaviours have been recorded and tested in restricted environments where the number of participants is relatively small (2 people: [19], 4 people: [2, 14]), though some studies have started to consider scenarios where more people (7-12) are involved [22, 13]. In most studies, the participants are seated or remain seated for most of the time, which inhibits the natural inclination of people to adjust their proximity and body orientation towards each other when they converse. In addition to the visual and and auditory senses, the ability to adjust one's inter-personal proximity has a significant influence on the perception of smells, touch or body temperature of the other person [11] and can indicate significant differences in the relationships between people. By addressing natural social behaviour in cases where people are standing we are able to capture a wider range of body language. This provides important cues about a person such as their personality, [22], and also their relationship with other people e.g. if they are in an intimate relationship [7].

Figure 1: Screen-shot from our data (best seen in colour). Some example F-formations are circled. The associate on the right side of an F-formation is indicated by an arrow.

In this work we focus on freely formed and congregated groups of individuals who are meeting purely to exchange information or foster existing relationships (see Fig. 1). The gathering is such that most of the approximately 50 participants are acquainted with about half of the people at the event, This differs from the scenarios usually used to detect people who are walking together as the scenes are more likely to be filled by mostly unacquainted groups [23, 9, 1]. Unacquainted groups will interact with each other through avoidance strategies but this is clearly a different form of communication than actually having a conversation.

As the number of people in a scene increases, the likelihood of conversations being restricted to a single floor reduces significantly [6] and people tend to split off into separate groups to converse more directly with one another. Under such circumstances, the first challenge in understanding the relationship between group members is to be able to identify them. An initial step towards this, which has received relatively little attention and is the focus of this paper, is the detection of *focused encounters*. By approaching the task of detecting focused encounters, we provide a solid framework for measuring more qualitative behaviour such as aggression [21], or dominance [14]. Focused encounters

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
WOODSTOCK '97 El Paso, Texas USA
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$10.00.

were first defined by Goffman [10] as a collection of people who are gathered together such that a shared space is maintained between them within which a conversational exchange can occur. They differ from *unfocused encounters* [10] which refer to the way people move to ensure they do not bump into each other, or greeting someone when walking past them, for example. An *F-formation* is considered as a specific instance of a focused encounter [5], which we will describe in more detail in the next section. Unfocused encounters and focused encounters can appear very similar since proximity and orientation cues can be used in both cases. However, clearly interactions during focused encounters can have a much more substantive meaning as all the participants co-operate to sustain a continued period of information exchange. Distinguishing between them allows us to identify differing levels of potential influence.

Detecting F-formations automatically may appear to be a relatively simple task which can be easily extracted from who is 'standing with whom' (as suggested by Yu et al. [20]), but one quickly finds that in more crowded situations where perhaps, the more interesting behavioural phenomena are to be found, proximity can be significantly affected by the layout of a room (e.g. position of furniture) or how crowded it is [5]. Body orientation also plays a role as it provides a prior on the direction of attention of a person but can again be influenced by the same external factors. One can also observe frequent situations at social gatherings where someone tries to join a group but is left standing on the periphery, trying to listen or join in on the conversation. If no one in the group lets him, he clearly has a different status to the full participants in the conversation. Therefore, there are different aspects to being involved in a focused encounter, which goes beyond a person's willingness (through their body behaviour) to be part of it.

The contributions of this paper are twofold. First, we address the new problem of identifying F-formations in crowded social scenes. The classification task is informed by prior work in psychology and social sciences and provides a principled framework from which more complex social behaviours can be identified. Second, we make improvements over the baseline method proposed by Yu et al. [20] by formulating the problem as one of identifying dominant sets [17]. A dominant set is a form of maximal clique that can be applied to edge weighted graphs so that the affinity between all nodes within it is higher than that between the internal nodes and those that are external to it. This differs from *modularity cut* [16] which is a global optimisation method where partitions of the graph are made according to the difference between the affinity of two nodes and the expected degrees of the vertices. Since participants of an F-formation require equal access to a shared space, clustering people within them can be considered more of a local rather than global optimisation task. Finally, we suggest a modification to Pavan and Pelillo's peeling off strategy, which is used to identify all the dominant sets in a graph, by introducing a stopping criterion which enables better detection of singletons.

For the remainder of this paper, we will discuss related work. Then we motivate the need to consider F-formations as dominant sets (Sec. 3). We describe the data and annotation process used to evaluate our methods in Sec. 4. In Sec. 5, we provide a brief description of the modularity cut algorithm to highlight the differences between it and our proposed clustering method (Sec. 6). In Sec. 7, we provide

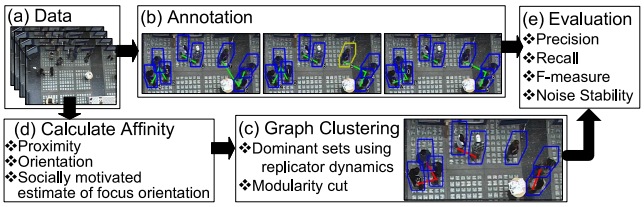

Figure 2: Flow diagram summarising our approach. Green lines:annotated F-formations, Yellow box: associate, Red lines: detections.

a description of how the affinity matrix was calculated using proximity and orientation information as well as our socially motivated estimate of visual focus, which is calculated from location information only. We present our experimental results in Sec. 8 and conclude in the final section. Our approach is summarised in Fig. 2.

2. RELATED WORK

Goffman [10] defined focused interaction as being "concerned with clusters of individuals who extend one another a special communication license and sustain a special type of mutual activity that can exclude others who are present in the situation"(p. 83). He went on to say that such a spatial and orientational arrangement "tends to be carefully maintained, maximizing the opportunity for participants to monitor one another's mutual perceivings"(p. 95). Ciolek and Kendon [5] took this idea further by defining a focused interaction more precisely as an F-formation

> "whenever two or more individuals in close proximity orient their bodies in such a way that each of them has an easy, direct and equal access to every other participant's transactional segment, and when they maintain such an arrangement, they can be said to create an F-formation" (p.243).

The transactional segment is the region in front of the body where limbs can reach easily, and hearing and sight is most effective. They defined the *F-formation system* to be the system of spatial and postural behaviours by which people create and sustain the shared interaction space between them. *Associates* of an F-formation were defined as people who may adjust their position relative to the F-formation system but are not included in its boundaries; they can come and go from an F-formation without the usual rituals that full participants of an F-formation would undergo. Examples of F-formations and associates are shown in Fig. 1.

In the computer science community, work on detecting groups has tended to focus more on finding people who are 'together' based on the persistence of their proximity and direction of motion [9, 23, 1, 18] during walking, which are readily measurable from the extracted trajectories. In these scenes, individuals walk with people they know and it is unlikely that they will be acquainted with other walking groups. Unfocused encounters, are therefore much more prominent. Focused encounters on the other hand, give a far richer semantic meaning to the interactions between people. To our knowledge, little work has tried to identify the key factors involved in being in a focused encounter.

Zen et al. have addressed some of the issues involved in approaching people to interact with them, within the framework of estimating personality traits from proxemic behaviour [22]. However, the identification of F-formations was not explicitly addressed.

The closest work to ours was done by Yu et al. [20], who proposed a system that could track and discover groups of interacting people. The *modularity cut* algorithm [16] was applied to identify these groups from automatically extracted trajectories. A problem with this work is the choice of data and experimental design. The 30 minute test data set was rather artificial as 23 people were asked to "mingle in a 3-group configuration" [20] (p.1468). Three groups were indeed identified but given the average number of people per group, it is unlikely that such large F-formations could be sustained [6]; typically, large gatherings tend to split or merge into smaller F-formations over time as the conversation changes. It is difficult to conclude what the 3 groups represented semantically, other than spatially separated locations. The same authors have subsequently used modularity cut to identify groups for classifying group activities in a prison scenario [3] where collective aggressive behaviour from people acting as 'gangs of inmates' are analysed. Also, they used only proximity information and not body orientation, when computing the affinity between detected people. Other work which is close to ours is that of Brdiczka et al.[2] who analysed speech activity to identify who was interacting with whom among 4 subjects. Their algorithm could automatically identify pre-scripted periods when different combinations of pairs of people were speaking or the whole group was speaking together. It is not clear if their algorithm could scale to larger group sizes and also whether using more realistic conversational data would have provided similar results.

3. GRAPH-THEORETIC DEFINITION OF F-FORMATIONS

The people in a scene can be represented as a graph $G = (V, E, w)$ with a set of vertices V, edges E, and a positive edge weight function w. In this case, the vertices or nodes are the people, E correspond to the set of connections between the people and w represents the affinity measured using some extracted features between each pair in the scene. We can express the relationship between all the nodes or people in the scene by a weighted affinity matrix $\mathbf{A}$ such that each of its elements $a_{ij} = w(i, j)$.

3.1 F-formations as High Modularity

Yu et al. [20] defined the task of identifying groups of people in terms of the recursive global partitioning of the graph by maximising the remaining modularity of the uncut edges. Modularity was proposed by Newman [16] as a metric for clustering social networks. The modularity between two nodes in a graph is represented in the corresponding element of *modularity matrix* $\mathbf{B}$ where each of its elements is defined

$$b_{ij} = a_{ij} - \frac{k_i k_j}{2m}. \quad (1)$$

a_{ij} is the affinity between node i and j, k_i and k_j are defined as the expected degrees of the vertices, indicating how 'connected' that particular node is to the rest of the network; $k_i = \sum_j a_{ij}$, and $2m$ is a normalisation term; $m = \frac{1}{2}\sum_{ij} a_{ij}$.

For a network with n nodes, which is divided into two groups by the indicator vector $\mathbf{s}$, each of its n elements are either labelled as $s_i = 1$ for all vertices in group 1 and $s_i = -1$ for those in group 2. The modularity of the entire network

$$Q = \frac{1}{4m}\sum_{ij} b_{ij} s_i s_j, \quad (2)$$

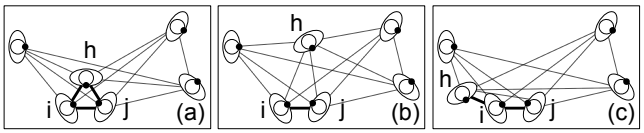

Figure 3: Example of how modularity would differ counter-intuitively. Thick lines represent high affinity while thin lines show low affinity. (a): h, i, and j are in an F-formation. (b): i and j are in an F-formation while h is far away from them. (c): i and j are in an F-formation while h is very close to i.

defines how well connected it is when cuts are made according to $\mathbf{s}$. The main assumption of modularity is that the affiliation between any pair of nodes is only significant if its value is greater than the expected affiliation of those two nodes with the rest of the network.

The modularity is well adapted to finding two-node maximal cliques but is not specifically designed for larger clique sizes, as can be demonstrated by the following example. Let us suppose that we have a network where nodes i, j and h are in an F-formation (Fig. 3(a)). In this case, b_{ij} would be smaller given the proximity of node h compared to if node h was not part of the F-formation (Fig. 3(b)). If h was standing closer to i than j (Fig. 3(c)), thus breaking the triadic F-formation, b_{ij} would be higher than if h was standing equally close to i and j. So in (Fig. 3(a)), b_{ij} is penalised because of the proximity of h to i and j whereas it makes more sense that the proximity of h should *support* the affinity of i and j.

3.2 F-formations as Dominant Sets

As already mentioned, an F-formation is defined as a group of people who have easy and equal access to the same shared space, around which they can communicate for a prolonged time. By its definition, therefore, mutual affinity between all of its members should be higher than the affinity between any of its members and those outside of it. In the case of an associate of an F-formation (e.g. someone who wishes to, but is unsuccessful at joining the group), they are clearly close to one or two members of the F-formation but may not have access to all of them. Based on these definitions, it seems clear that using modularity to identify F-formations is not precise enough for our task.

Pavan and Pelillo [17] proposed a different way of thinking about a cluster as a dominant set, which is a generalisation of maximal cliques to edge-weighted graphs. If we consider a subset S of the set of nodes V in graph G, the average weighted degree of a vertex $i \in S$ with respect to set S is

$$k_S(i) = \frac{1}{|S|}\sum_{j \in S} a_{ij}. \quad (3)$$

Note that in this definition of the degree of i, the value is strictly related to only a subset of the graph V. Ideally, S defines a semantically meaningful local context such as the F-formation consisting of persons i, j, and h in Fig. 3 (a). The relative affinity between node $j \notin S$ and i is

$$\phi_S(i, j) = a_{ij} - k_S(i), \quad (4)$$

and the weight of each i with respect to a set $S = R \cup \{i\}$ is defined recursively as

$$w_S(i) = \begin{cases} 1 & \text{if} |S| = 1 \\ \sum_{j \in R} \phi_R(j, i)\, w_R(j) & \text{otherwise} \end{cases} \quad (5)$$

and $\phi_{\{i\}}(i,j) = a_{ij}$. $w_S(i)$ measures the overall relative affinity between i and the rest of the vertices in S, weighted by the overall affinity of the vertices in R. Therefore $w_{\{ij\}}(i)$ and $w_{\{ij\}}(j)$ would not vary for any of the conditions in Fig. 3, while $w_{\{ijh\}}(h)$ would be highest for (a), lower for (c), and lowest for (b). This relationship between internal and external nodes of a dominant set S is defined formally using the following conditions;

$$w_S(i) > 0, \forall i \in S \tag{6}$$

$$w_{S\cup\{i\}}(i) < 0, \forall i \notin S. \tag{7}$$

Therefore, dominant sets describe very compact structures, which is ideally suited to represent F-formations of any size.

4. OUR DATA

The data we used (see Fig. 1) consists of real footage of over 50 people who met to present scientific work during a poster session. The focused encounters that are formed were all natural and unscripted, motivated only by each individual's real relationships with other people at the event. The data captures interactive behaviour between people who are closely acquainted (e.g. work colleagues or friends), as well as strangers. The event lasted for approximately 3 hours and was captured by video from a camera that was mounted approximately 15 m overhead. The distance and orientation of the camera helps to conserve the privacy of the participants while enabling the analysis of a crowded scene. To our knowledge, this is currently the largest data set which captures naturally formed focused encounters.

4.1 Annotating the Data

Images from the data were selected so that each one contained different F-formations. In total, 82 images were selected for annotation, containing $\sim$ 1700 people. Selection was made based on leaving at least 10s between images and that no consecutively selected images contained the same formations of people. We also tried to maximise on the crowdedness and ambiguity of associates in the scenes. 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. 24 annotators from different international cultural backgrounds volunteered to label the data. The annotators were grouped into 3-person subgroups to label the same data so that variability in the labelling could be taken into account during evaluation. After being given appropriate definitions, annotators were asked to identify F-formations and their associates from static 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. The annotators labelled 10 - 11 images each, with on average $\sim$ 170 instances of people to label. After the annotation task, the annotators were asked to provide free comments about what they found difficult or easy about the task. Cues that they reported to use for determining who was a member or associate of an F-formation included gaze direction, proximity and body orientation.

4.2 Annotation Analysis

We analysed the annotations to see how many sizes of F-formation there were and how frequently each size occurred based on whether there was full agreement between the annotators about all members of an F-formation, or when the union of all overlapping groups was considered. The frequency of occurrence of each F-formation size is shown in Table 1. 340 singletons were labelled, with 450 full agreement F-formations and 599 when the union was used. Using both cases, dyads occurred most frequently. The form of the distribution aligns with those found by Ciolek who studied how people gathered in public areas [4]. The level of annotator agreement was calculated using the F-measure per F-formation (or singleton) for each pairwise combination of annotators who labelled the same data. The mean of these values was taken for each triad of annotators and then finally the average for all groups of annotators was calculated. The mean average F-measure was 94.74% so the annotator agreement was high.

F-formation Size	1	2	3	4	5	6	Total *
Full Agreement	340	378	69	2	1	0	450
Union	340	452	136	8	1	2	599

Table 1: Distribution of labelled F-formation sizes using full agreement or the union of labelled members. *Not including singletons.

5. MODULARITY CUT

To highlight the difference between our proposed method and modularity cut, we will provide a brief introduction here. Further details can be found in Newman's original paper [16]. Modularity cut was originally developed for social network analysis and discovers groups in social networks, where the size of the communities being tested on are large and consequently, the discovered communities are also of quite large. Given this application domain, maximal cliques or dominant sets are less likely to occur and for discovering social groups based on people with common hobbies or interests, the density of connections between community members is not as important as how connected each person is to everyone else in the network. For the task of identifying spatially separated groups, Yu et al. [20, 3] have shown the effectiveness of using modularity cut. However, as we have demonstrated in Sec. 3, modularity may not be the best representation of F-formations.

From Eq. 2 in Sec. 3, we can see that by carefully selecting $\mathbf{s}$, it is possible to maximise the modularity Q by summing over the higher-valued elements of the modularity matrix $\mathbf{B}$. Since Eq. 2 can be conveniently arranged so that

$$Q = \frac{1}{4m}\mathbf{s}^T\mathbf{B}\mathbf{s}, \tag{8}$$

performing Eigen decomposition on $\mathbf{B}$ and choosing appropriate values for the elements of $\mathbf{s}$ such that the largest or most positive eigenvalues are given the most weight, will lead to a maximisation of Q. In our case, however, the elements of $\mathbf{s}$ are restricted to be either ± 1 and finding the optimal solution for $\mathbf{s}$ is likely to be NP-hard. In practice, a good approximate solution can be found by taking the principal eigenvector of $\mathbf{B}$, $\mathbf{u_1}$, and setting the sign of each element of $\mathbf{s}$ to match those of $\mathbf{u_1}$. Note that the degrees, k_i, are always calculated relative to all the other vertices in the network so the graph is partitioned using a global context. In practice, this can have negative consequences for F-formation estimation, particularly when we consider how people typically arrange themselves spatially in a room. This can be influenced by both the furniture and function of the occasion. In our case, people typically tend to cluster around the edges of the space either to watch others passing through,

or to present a poster. This means that the people tend to distribute in clusters of rings around the space. Under such circumstances, a partition can occur that splits apart an F-formation near the centre of the scene just because the scene is equally crowded on both sides. The modularity is still maximised as the people in this central F-formation are, on average, the furthest away from the rest of the nodes in the graph so that cutting most of the edges still leads to a small reduction in modularity. This particular example indicates well the problem of treating F-formation detection as a purely global optimisation task.

5.1 Kernighan-Lin (KL) Refinement

Since the principal eigenvector can only provide an approximation for the division, a step inspired by the Kernighan-Lin algorithm [16] can be applied to try to maximise Q further. This involves taking each node in turn, moving it to the other group and then recalculating Q. The node change which results in the highest Q (even if this leads to a negative change in modularity) is kept and then the process is repeated until all nodes have been swapped. If the final modularity leads to an overall improvement, then the partitioned refinement is kept. This step has been used widely for network analysis with significant improvements [16].

5.2 Further Subdivisions of the Network

Recursive division of the network is carried out by considering an $n_G \times c$ matrix $\mathbf{S}$ which serves to indicate with a one when a node n belongs in community c and zero otherwise. G is the sub-community of the network (with n_G nodes), that is being considered for further subdivision. By measuring the change in modularity before and after the subdivision of the network, we can decide to keep the split based if it leads to a decrease in modularity.

$$\begin{aligned} \Delta Q &= \frac{1}{2m}\sum_{i,j\in G}\sum_{k=1}^{c} B_{ij}S_{ik}S_{jk} - \sum_{i,j\in G} B_{ij} \\ &= \frac{1}{2m}\mathrm{Tr}(\mathbf{S}^T\mathbf{B}^{(G)}\mathbf{S}). \end{aligned} \tag{9}$$

$\mathbf{B}^{(G)}$ can be considered as the sub-matrix of $\mathbf{B}$ for the subgraph G. By re-arranging ΔQ to have the same form as Eq. 8, we can conveniently re-apply the same eigenvector-based selection process to subdivide the communities. To achieve this form however $\mathbf{B}^{(G)}$ is defined differently to $\mathbf{B}$:

$$B^G_{ij} = B_{ij} - \delta_{ij}\sum_{l\in G} B_{il}. \tag{10}$$

6. IDENTIFYING DOMINANT SETS WITH REPLICATOR DYNAMICS

As described in Sec. 5, one way of clustering the graph is to recursively partition it using spectral techniques such as eigen decomposition. However, finding dominant sets could be considered better suited to a local optimisation problem. In doing so, we also ensure that the resolution of the clique sizes are not affected by the number of positive links and the cardinality of the graph [8]. Pavan and Pelillo [17] showed that the notion of a cluster and its relationship to dominant sets were mathematically equivalent by formulating the optimisation problem as a standard quadratic programme where

$$f(x) = \mathbf{x}^T\mathbf{A}\mathbf{x} \tag{11}$$

is maximised subject to the constraint that $\mathbf{x}$ lies on the standard simplex $\Delta = \{x \in \mathbb{R}^n : \mathbf{x} \geq 0 \text{ and } \sum_n x_n = 1\}$. $\mathbf{A}$ in this case is still defined as the affinity matrix of the graph. However, $\mathbf{x}$ is slightly different from the indicator vector $\mathbf{s}$ that was used to maximise the modularity in Eq. 8. Here, $\mathbf{x}$ is defined as the weighted characteristic vector and when defined in terms of the subset of vertices S,

$$x^S_i = \begin{cases} \frac{w_S(i)}{W(S)} & \text{, if } i \in S \\ 0 & \text{, otherwise} \end{cases}, \tag{12}$$

where $W(S) = \sum_{i\in S} w_S(i)$ is the total weight, which must be greater than 0. Pavan and Pelillo [17] proved that by using this definition of $\mathbf{x}$, the maximisation of the objective function is the same as finding dominant sets. Further details of this proof can be found in [17].

To find a local solution of the objective function f, a method taken from evolutionary game theory, called replicator dynamics was used. The first-order replicator equations are defined as

$$x_i(t+1) = x_i(t)\frac{(\mathbf{A}\mathbf{x}(t))_i}{\mathbf{x}(t)^T\mathbf{A}\mathbf{x}(t)} \tag{13}$$

and are applied to all nodes in the network in turn. Since $\mathbf{A}$ is symmetric, the replicator equations provide a strictly increasing update to the characteristic vector $\mathbf{x}$, which converges upon the local solution of f. By taking the support or non-zero indices of the final $\mathbf{x}$, we identify the elements of the graph that are a dominant set. That is, the solution of the replicator equations converges exactly on a characteristic vector that conforms exactly to conditions 6 and 7. In practice, $\mathbf{x}$ is initialised with uniform weights, which corresponds to the centroid of the standard simplex.

6.1 Further Subdivisions of the Network

Since the replicator equations only converges on the most dominant set of a particular graph, an effective way of identifying further clusters in the network is to apply a peeling strategy [17]. This involves identifying a dominant set using Eq. 13, removing the corresponding nodes, and then re-applying the replicator equations to the remaining subgraph. In practice, the elements of the characteristic vector rarely converged to exactly 0 so a threshold was used to identify numbers that were extremely close. The same threshold was also used to ensure that if the value of the maximised objective function was too small, we considered that all the remaining nodes in the graph were singletons.

6.2 Local vs Global Context

There is one disadvantage of using a peeling off strategy to determine dominant sets. As more dominant sets are removed from the original graph, fewer and fewer nodes remain until it is likely that those that are not clustered yet are more likely to be singletons than in an F-formation. We modify the peeling strategy suggested by Pavan and Pelillo [17] by introducing a more principled stopping criterion, which takes into account the global context of the complete graph. Given the nature of dominant sets, we know that conditions 6 and 7 must hold. Using the peeling off strategy, this is certainly the case for the dominant set that has been identified within the sub-graph. However, if we compare $w_{S\cup\{i\}}(i)$, for all $i \notin V$, where V is the set of all nodes in the entire graph, we also ensure that the clustering makes sense within the global context and the need for assigning a slightly arbitrary threshold as a stopping criterion is removed.

7. BUILDING THE AFFINITY MATRIX

7.1 Proximity

Perhaps the most obvious way of measuring the affinity between people is to use their relative proximity, as proposed by Yu et al.[20]. The symmetric distance function between person i and j is defined as

$$A_{ij}^{prox} = -e^{\frac{d_{ij}}{2\sigma^2}} \tag{14}$$

where d_{ij} is the Euclidean distance in the ground plane between person i and j and σ is the variance of the function. Given the nature of our data, σ was set to approximately 2 metres; the commonly accepted distance at which focused encounters occur [12]. Note that unlike Yu et al., our affinity matrix is generated from a single static image.

7.2 Proximity and Orientation

Using proximity alone can lead to errors in the F-formation estimates as the space becomes more crowded. Fortunately, the body orientation can help to identify the shared space between all the F-formation members. This is particularly relevant for cases where associates exist as they may be oriented towards an F-formation but those closest to the associate may have their backs to them. To address this case, A_{ij} is defined by taking the worst case in both directions;

$$A_{ij}^{ori} = \underset{\mathbf{q}}{\text{argmin}}\left(\begin{cases} e^{-\frac{d_{\mathbf{q}}}{2\sigma^2}} & -\frac{\pi}{2} \geq \theta_{\mathbf{q_1}} - \alpha_{\mathbf{q}} \geq \frac{\pi}{2} \\ 0 & \text{otherwise} \end{cases} \right), \tag{15}$$

$\forall \mathbf{q} \in \{(i,j),(j,i)\}$, where θ_i is the angle of body orientation of person i, $\mathbf{q_1}$ is the first element of q, and α_{ij} is the angle of the vector from i to j. An asymmetric or directed version of A_{ij} can be considered by taking the actual values of the affinity rather than the minimum in both directions and using an extension of modularity that can deal with a directed affinity matrix[15]. Though our experiments with a directed graph performed better than the basic modularity cut, using the undirected graph led to better performance.

7.3 Socially Motivated Estimate of Focus Orientation

Compared to person position, body orientation can be particularly difficult to extract robustly in crowded environments. To bypass the need to extract body orientation information as well, we propose a *socially motivated estimate of focus orientation* (SMEFO) γ based on the relative distance between participants in the scene. Given a person in the scene, we expect that they are most likely to be oriented towards the people they are motivated to converse with. If the motivation to talk to others is correlated with proximity, people will tend to be much closer to those that they wish to talk to. The SMEFO is defined as

$$\gamma_i = \arccos(\alpha(p_i, f_i)), \tag{16}$$

where p_i is the location of person i and $\alpha(p_i, f_i)$ is the angle of the vector from person i to their estimated centre of focus

$$f_i = \frac{1}{k_i} \sum_j p_j A_{ij}^{prox}. \tag{17}$$

θ in Eq. 15 can then be replaced by γ. Note that k_i here is the degree of i for the entire graph.

8. EXPERIMENTS

To evaluate our methods, the precision, recall and f-measures were calculated for each F-formation and singleton, labelled by the annotators. An F-formation in the annotations was considered to be the union of all labelled groups that overlapped. The groupings were scored for each labelled F-formation by comparing each person in it to the corresponding detected group. Any overlapping nodes between the labelled and detected groups was given a score of $\frac{1}{3}$ per annotation, depending on whether the detection was a true positive, false positive, or false negative. As there were three annotators per image frame, given a labelled group, each person in it could be scored from 0 up to 1, depending on how many annotators agreed that the person was in the F-formation. Then, the set of annotations was compared against the detected F-formation and accumulated to get the precision, recall and f-measure of that particular group. We report the average of each of these measures over all the groups. For the purposes of evaluation the set of all possible groups consisted of all the F-formations and also the singletons. The position and orientation of people were manually labelled in the images, which were then used for building the affinity matrices in Sec. 7. The positions were projected onto the ground plane using orthogonal projection.

First, we present our results using a factorial combination of the different features and methods that were described earlier, shown in Table 2. The features used were proximity (**P**: Eq. 14), proximity and the manually labelled orientation (**P+O** : Eq. 14, 15), and proximity and the SMEFO feature (**P+F**: Eq. 14–16). The methods we used were the dominant set identification using replicator dynamics (**DS**: Eq. 13) with local context where a threshold on the value of the objective function was used (**LC**), and the global context where a stopping criterion was defined based on when $w_S(i) > 0$ for any i not in the identified dominant set S (**GC**), as described in Sec. 6. We compare our results with the method of Yu et al. [20], that used modularity cut and proximity to create the affinity matrix (**MC**: Sec. 5).

For space reasons, we only show the best most automated method of extracting features in combination with modularity cut, which used the proximity and SMEFO features with modularity cut and Kernighan-Lin refinement (**MC+KL**), though the performance in all cases was improved over [20]. In addition to the performance of different methods, we also provide a baseline based on if all people in the data were labelled as singletons. Note that the performance of this baseline is quite high because most of the high performance is distributed in the set of annotated singletons. Note also the high precision is biased towards the naturally high precision of detecting singletons in the data.

Table 2 shows that all our proposed methods out-performs that the baseline [20]. Both fully manual methods with our proposed dominant set technique (**P+O, DS**) performed the best but was comparable to using **(P+O,MC)**. If we examine the performance of the methods when not using the manually labelled body orientation, a clearer picture emerges about the effectiveness of both methods. It is particularly interesting to note that using the position information only (**P, DS+GC**), the dominant set method outperforms all methods that use the SMEFO feature. When considering proximity features only, both **DS** methods outperform the corresponding modularity cut method by over 8.2% in absolute terms. Compared to (**P+F,MC +KL**),

Cues	Methods	Prec	Recall	F-Meas
P	**DS+LC**	80.24	95.72	84.85
P	**DS+GC**	87.29	91.79	**86.83**
P	**MC**	68.72	97.43	76.57*
P+O	**DS+LC**	90.36	96.63	91.87
P+O	**DS+GC**	91.94	95.57	**92.24**
P+O	**MC**	89.50	98.40	92.02
P+F	**DS+LC**	83.08	94.25	85.85
P+F	**DS+GC**	85.40	93.26	**86.50**
P+F	**MC+KL**	77.06	96.28	82.32
All Singletons		95.63	66.27	75.57

Table 2: Results summary in terms of the mean precision, recall and F-measure per F-formation. The boldened values show the highest average F-measure per F-formation for each feature combination. The last row shows the performance if everyone was labelled as singletons. *Baseline from Yu et al. [20].

the absolute improvement of (**P, DS+GC**) is 4.5%, despite not taking into account any orientation information.

The SMEFO feature was proposed as another way of addressing the problem of associates of F-formations who are often close to full participants but may not have equal access to all of them. Using SMEFO, it was possible to identify some cases where someone clearly would not be able to converse with the other person because there would be someone else in between. When using dominant sets to identify the F-formations, this idea of equal mutual access between all the participants is represented more explicitly, which may explain why the performance is better.

The similar performance of both dominant sets and modularity cut, when using both manually annotated location and orientation information, suggests that body orientation is a strong cue for detecting F-formations. However, while the increase in performance when using fully labelled data compared to location information only with modularity cut is 15.5%, using **DS**, the performance improvement is much less, at 5.41%. This suggests that the importance of body orientation may be less important than ensuring equal and high affinity between all members of the F-formation. It also questions whether using more direct estimates of body orientation (i.e. from the imagery data) would necessarily provide a significant improvement for F-formation estimation compared to using position information alone. When comparing the singleton baseline to all other methods, we see that using the basic modularity cut algorithm proposed by Yu et al. [20] only just out performs it, indicating how poor modularity cut is for detecting F-formations. We also found that for the dominant set method, the **GC** method consistently out-performed the corresponding **LC** method, which demonstrate the effectiveness of our improvement to the stopping criterion of the peeling back strategy.

8.1 Simulating Tracking Errors

So far, our experiments have shown F-formation detection using manual annotations of the position and body orientation information. Using the SMEFO feature, while the body orientation becomes an estimate, the positions are still manually extracted. To test the stability of our method compared to the baseline methods, we applied increasing levels of Gaussian noise to the manually labelled positions. For cases where we tested the clustering methods using manually annotated body orientation information, we did not add noise to the orientations as adding noise to the positions, given the body angle, would already affect the affinity sufficiently. It also provided a more consistent framework from which to compare the different features and methods. The results are summarised in Figure 4.

We noticed when analysing the breakdown of the performance across different group sizes, that a significant increase in the performance of the dominant set method over modularity cut was due to significantly better estimation of singletons. Since these are not really considered to be F-formations, we provide the performance based on all F-formations that contain at least 2 people. This is indicated as a thin black line in Figure 4, while the thick black line shows the overall performance for all cluster sizes. We also noticed when applying noise to the positions, that the F-formation performance tended to saturate as the noise level increased. On closer inspection we found that there was a tendency for all nodes to be labelled as singletons as they became more spread out. To illustrate this, we also show a dashed line in Figure 4, which represents the performance when all nodes are labelled as singletons, and a dot-dashed line which shows the performance for the same method evaluated without the singletons. As expected, when the labelled singletons are excluded from the calculation, the performance drops significantly from 75% to 62%.

Note that using modularity cut, the performance increases significantly when singleton detection is not evaluated, which suggests that modularity cut is better at detecting larger group sizes. In almost all cases when singletons are not evaluated, the dominant set method performs worse. In terms of noise stability, the deterioration in performance occurs much sooner for modularity cut, compared to the dominant set methods. For the SMEFO feature, it provides more stability during noisy conditions for the modularity cut method, while improvements for the dominant set method is less significant. Using the manual body orientation labels, however, appear to provide much more stability for the **DS** method, suggesting that estimates of body orientation may still be useful when the tracking estimates are very noisy.

9. CONCLUSIONS AND FUTURE WORK

In this paper, a new approach for detecting F-formations was presented by formulating the problem as one of finding dominant sets. Compared to modularity cut [20], significant and more stable performance improvements were observed. Our SMEFO feature also provided better performance, showing that some estimate of body orientation can help to improve the F-formation estimation performance but having accurate position information is more likely to provide better estimates. Since the method currently does not use automated person detection or body orientation, further investigations are needed to see the effect on F-formation estimation. Our test data is currently limited to scenarios where everyone is standing. In public spaces, a variety of sitting and standing behaviours occur depending on possible seating areas or other furniture. This was represented to some extent in our current data with high circular coffee tables but could be extended to more extreme cases. Increased furniture in the space can lead to spatial deformations of the F-formations that occur more in standing scenarios. So dominant sets constructed from instantaneous spatial cues alone will have limited success. Examining co-ordinated conversational movement in video may help to mitigate this problem.

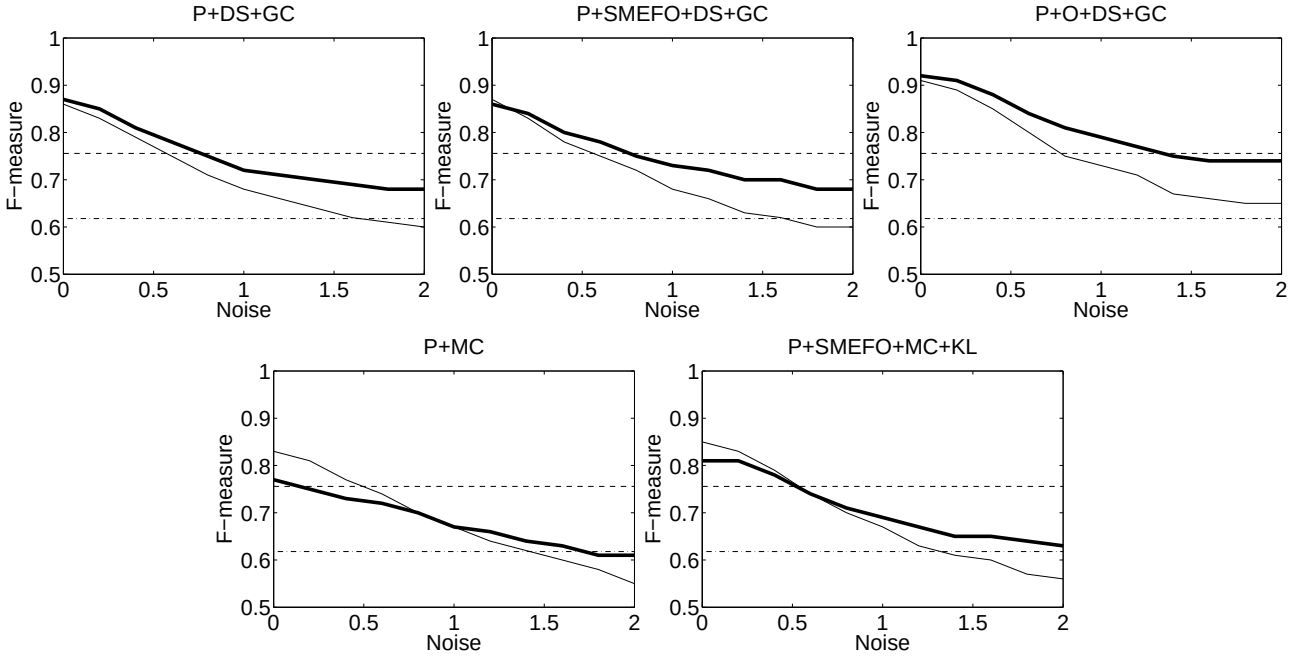

Figure 4: Comparison of results when Gaussian noise was added to the people's positions. The noise is expressed in terms of the proportion of the approximate width of a person. Thick black line shows the mean f-measure evaluating over all F-formation sizes while the thin black line indicates the performance when labelled singletons were excluded from the evaluation. Horizontal lines indicate the f-measure if every node was labelled as a singleton (dashed lines) and when labelled singletons were not evaluated (dot-dashed lines).

Acknowledgements

This research was supported by a Marie Curie Research Training Network fellowship in the project "AnaSID" (PIEF-GA-2009-255609). We thank Jagan Varadarajan, Bastien Crettol, and Valérie Devanthéry for their help during the capture and processing of the data, and Gwenn Englebienne for his template-based orthogonal projection code.

10. REFERENCES

[1] A.French, A. Naeem, E. Dryden, and T. Pridmore. Using social effects to guide tracking in complex scenes. In *AVSS*, pages 212–217, 2007.

[2] O. Brdiczka, J. Maisonnasse, and P. Reignier. Automatic detection of interaction groups. In *Proceedings of the 7th international conference on Multimodal interfaces*, ICMI, pages 32–36, 2005.

[3] M.-C. Chang, N. Krahnstoever, S. Lim, and T. Yu. Group level activity recognition in crowded environments across multiple cameras. *AVSS*, pages 56–63, 2010.

[4] T. M. Ciolek. Location of static gatherings in pedestrian areas: an exploratory study. *Man-Environment Systems*, 7:41–54, 1977.

[5] T. M. Ciolek and A. Kendon. Environment and the spatial arrangement of conversational encounters. *Sociological Inquiry*, 50(3-4):237–271, 1980.

[6] R. Dunbar, N. Duncan, and D. Nettle. Size and structure of freely forming conversational groups. *Human Nature*, 6(1):67–78, 1995.

[7] J. J. Edney and N. L. Jordan-Edney. Territorial spacing on a beach. *Sociometry*, 37(1):92–104, March 1974.

[8] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36, 2007.

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

[10] E. Goffman. *Behavior in Public Places: Notes on the Social Organization of Gatherings*. Free Press, 1966.

[11] E. T. Hall. *The silent language*. Doubleday, Garden City, N. Y., 1959.

[12] E. T. Hall. *The Hidden Dimension*. Doubleday, Garden City, N.Y. :, [1st ed.] edition, 1966.

[13] H. Hung and G. Chittaranjan. The wolf corpus: Exploring group behaviour in a competitive role-playing game. In *ACM Multimedia*, 10 2010.

[14] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech and Language Processing*, 17(3):501–513, 2009.

[15] E. A. Leicht and M. E. J. Newman. Community structure in directed networks. *Phys. Rev. Lett.*, 100:118703, 2008.

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

[17] M. Pavan and M. Pelillo. Dominant sets and pairwise clustering. *IEEE Transactions on PAMI*, 29, 2007.

[18] S. Pellegrini, A. Ess, K. Schindler, and L. J. V. Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV'09*, pages 261–268, 2009.

[19] R. Ranganath, D. Jurafsky, and D. McFarland. It's not you, it's me: Detecting flirting and its misperception in speed-dates. In *EMNLP'09*, pages 334–342, 2009.

[20] T. Yu, S. Lim, K. A. Patwardhan, and N. Krahnstoever. Monitoring, recognizing and discovering social networks. In *CVPR*, 2009.

[21] W. Zajdel, J. Krijnders, T. Andringa, and D. Gavrila. Cassandra: audio-video sensor fusion for aggression detection. *AVSS*, 0:200–205, 2007.

[22] G. Zen, B. Lepri, E. Ricci, and O. Lanz. Space speaks: towards socially and personality aware visual surveillance. In *1st ACM international workshop on Multimodal pervasive video analysis*, pages 37–42, 2010.

[23] W. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, September 2009.