A Game-Theoretic Probabilistic Approach for Detecting Conversational Groups

Sebastiano Vascon[1], Eyasu Zemene Mequanint[2], Marco Cristani[1,3], Hayley Hung[4] ⋆, Marcello Pelillo[2], and Vittorio Murino[1,3]

[1] Dept. of Pattern Analysis & Computer Vision (PAVIS), Istituto Italiano di Tecnologia, Genova, Italy
[2] Dept. of Environmental Sciences, Informatics and Statistics, University Ca' Foscari of Venice, Italy
[3] Dept. of Computer Science, University of Verona, Italy
[4] Faculty of Electrical Engineering, Mathematics and Computer Science, Technical University of Delft, Netherlands

Abstract. A standing conversational group (also known as F-formation) occurs when two or more people sustain a social interaction, such as chatting at a cocktail party. Detecting such interactions in images or videos is of fundamental importance in many contexts, like surveillance, social signal processing, social robotics or activity classification. This paper presents an approach to this problem by modeling the socio-psychological concept of an F-formation and the biological constraints of social attention. Essentially, an F-formation defines some constraints on how subjects have to be mutually located and oriented while the biological constraints defines the plausible zone in which persons can interact. We develop a game-theoretic framework embedding these constraints, which is supported by a statistical modeling of the uncertainty associated with the position and orientation of people. First, we use a novel representation of the affinity between pairs of people expressed as a distance between distributions over the most plausible oriented region of attention.Additionally, we integrate temporal information over multiple frames to smooth noisy head orientation and pose estimates, solve ambiguous situations and establish a more precise social context. We do this in a principled way by using recent notions from multi-payoff evolutionary game theory. Experiments on several benchmark datasets consistently show the superiority of the proposed approach over state of the art and its robustness under severe noise conditions.

1 Introduction

After decades of research on the automated modeling of individuals, the computer vision community has recently started focusing on the new problem of analyzing groups [1,2,3,4,5,6,7,8,9,10]. In this paper, we focus on standing conversational groups, also known as *F-formations* [11], that is, groups of people who spontaneously decide to be in each other's immediate presence to converse with each and every member of that group. Standing conversational groups are of primary importance in many contexts, from video surveillance [7] to social signal processing [2,6,4,1], from multimedia [3] to social robotics [12] and activity recognition, as we will discuss extensively in Sec. 2. Many studies have been carried out by social psychologists to understand how people behave in public. By exploiting the theory behind these findings, we propose novel and

⋆ Author has been partially supported by the European Commission under contract number FP7-ICT-600877 (SPENCER) and is affiliated with the Delft Data Science consortium.

(a) (b) (c) (d)

Fig. 1. Standing conversational groups: a) in black, graphical depiction of overlapping space within an F-formation: the o-space; b) a poster session in a conference, where different groupings are visible; c) circular F-formation; d) a typical surveillance setting where camera is located at 2.5-3 meters from the floor, for which detecting groups is challenging.

more socio-psychologically principled ways of designing methods for automatically analyzing human behavior. For example, Hall [13] proposed that relationships and levels of interactions could be inferred by considering different social distances. Goffman [14] observed that group interactions can be categorized into those that are 'focused' and those that are 'unfocused'. Focused interactions concern the gathering of people to participate in an activity where there is a common focus, such as playing and watching a football match, conversing, or marching in a band. Unfocused encounters involves light interactions such as avoiding people on a street, briefly greeting a colleague while passing them in the corridor, or indicating to let someone pass when boarding a train.

Within the class of focused encounters, the F-formation is a specific type of group interaction which requires more attention from our senses. Specifically, an F-formation arises "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" [15, p. 243]. Some example of F-formations from real-world images are illustrated in Fig. 1a. There can be different F-formations as shown in Fig. 2a-e. In the case of two participants, typical F-formation arrangements are vis-a-vis, L-shape, and side-by-side. From an F-formation, three social spaces emerge: the o-space, the p-space and the r-space. The most important part is the o-space (see Fig. 2), a convex empty space surrounded by the people involved in a social interaction, in which every participant looks inward, and no external people are allowed. The p-space is a narrow strip that surrounds the o-space, and that contains the bodies of the conversing people, while the r-space is the area beyond the p-space.

Our goal in this paper is to develop a robust approach to automatically detect F-formations from images and videos employing a single monocular camera. As input, the approach requires the position of the persons in the scene on the ground plane as well as their body orientation, although in most cases, head orientation is more readily captured, even under heavy occlusion. These cues are easily obtainable nowadays, even if they are not estimated very accurately, and many approaches are devoted to these goals [16,17,4]. A recent experimental work of Setti et al. [18] shows that substantial improvement in the performance of F-formation detection algorithms can be achieved by combining a probabilistic approach such as the one developed in [7] and graph-based clustering methods [6]. Motivated by these findings, we develop a new, robust,

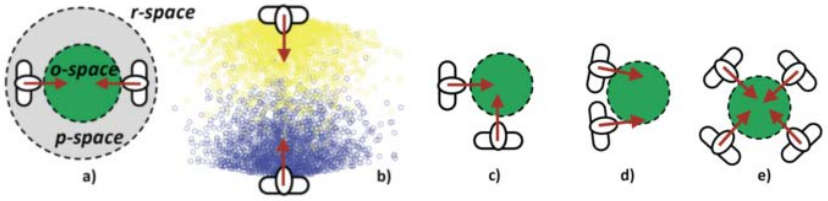


Fig. 2. F-formations; a) components of an F-formation: o-space, p-space, r-space; in this case, a face-to-face F-formation is sketched; b) modeling the frustum of attention by particles: in the intersection stays the o-space; c) L-shape F-formation; d) side-by-side F-formation; e) circular F-formation.

psychologically-principled approach which combines in a natural way the modeling of the uncertainty in the position and orientation of a subject and a game-theoretic clustering approach which allows one to extract coherent groups in edge-weighted graphs, digraphs and hypergraphs [19,20]. The game-theoretic setting provides a conceptual framework which allows also us to integrate temporal information in a principled way, in an attempt to reliably extract groups in video sequences under severe tracking noise. This is done by using a recent approach to integrate multiple payoff functions in an evolutionary game-theoretic setting [21].

Our approach is a substantial contribution for the computer vision community: so far, grouping behaviors have been analyzed mainly in dynamic situation via tracking, exploiting the oriented velocity as a primary cue, for example by associating individuals' tracklets [22,23,24,25,26,27,28,29,30]. In our case, F-formation are manifested primarily when people are still, so that a finer yet robust analysis is required.

To test the effectiveness of the proposed approach, we performed extensive experiments over five different datasets, each of which represents a particular scenario. In particular, we used a synthetic dataset [7], the Coffee Break dataset [7], the GDet dataset [7], the Idiap Poster data dataset [6], and the Cocktail Party [5] dataset. We also carried out systematic noise resistance experiments to fully investigate the stability of our method. The results consistently show the superiority of the proposed approach over the state of the art.

The rest of the paper is organized as follows. A detailed review of the literature on group detection approaches is presented in Section 2. Our approach is detailed in Section 3. In Section 4 we describe the game-theoretic clustering approach we use to extract F-formations and its extension to multiple affinity matrices. Finally, Section 5 presents the experimental results and Section 6 concludes the paper.

2 Literature review

2.1 Groups

During multi-party activities, we expect that there is a different underlying structure that governs the behavior of groups compared to individuals acting independently. For example, there has has been considerable prior work on estimating group activities by

modeling behavior at the individual as well as group level [8,9,10]. However, unlike works that treat all group structures equivalently, our premise is that there are fundamental semantic differences in what this prior work has considered to be a 'group' and what we refer (from the social psychological literature) as an 'F-formation' [11]. These prior definitions of a group of people assume that they are necessarily close together because they are for example, forming a queue, watching a football match, crossing the road together, or asked to mingle in a specific location. Some of these principles informed early socially-motivated methods of people tracking [31] by the social force model [32], that originated from pedestrian simulation research.

In more semantically meaningful social cases, one can attribute meaning to groupings based on some form of acquaintanceship, such as for detecting when people are traveling together [24] or when people are conversing in a lecture hall [2]. However in free standing scenarios, when people come together physically in order to make conversation, a specific, unspoken, and mutual agreement is made between all those involved that they wish to converse for some extended but finite period of time. Such an interaction requires a focusing of the senses, compared to the other group behaviors which can rely more on peripheral and unfocused sensing.

Importantly, the region in front of the body in which limbs can reach easily, and hearing and sight is most effective was defined as the *transactional segment*. A necessary condition of the F-formation was that the transactional segments of all members of an F-formation should overlap. Such a region can be considered an individual's frustum of social attention.

2.2 Exploiting visual attention

Considering this idea of frustum of attention, computer vision researchers have considered how the head pose can be used as a proxy for visual attention [33]. For visually led tasks such as looking at adverts [33], considering the visual attentional mechanisms is useful. However, when considering social contexts, the concept of social attention is a relatively new domain in the social sciences [34]. More specifically, head pose is actually equally if not more perceptually salient as a cue for gaze direction in humans [34, Ch. 6]. Moreover Kendon studied the role of gaze direction during conversational interactions suggesting that it functions as a cue for turn-taking, holding, or yielding [35]. Jovanovic and Op den Akker also found that addressees could be identified using gazing cues [36], while Duncan found that speakers attracted the gaze of listeners [37] during conversations. Ba and Odobez [38] exploited findings in primate social behavior by modeling plausible eye-in-head positions for gaze estimation to estimate the visual focus of attention of participants during meetings using only head pose while Subramanian et al. [39] used both gaze and head pose to estimate social attention in meetings.

2.3 Conversational groups detection

For the specific task of detecting F-formations, different approaches have been proposed. Groh et al. [1] proposed to use the relative shoulder orientations and distances (using markers attached to the shoulders) between each pair of people as a feature vector for training a binary classification task. Cristani et al. [7] proposed to solve the task

using a Hough voting strategy which accumulated a density estimating the location of the o-space. Concurrently, Hung and Kröse [6] proposed to consider an F-formation as a dominant-set cluster [40] of an edge-weighted graph where each node in the graph was a person, and the edges between them measures the affinity between pairs.

Later these two approaches were compared by Setti et al. [18] to investigate the strengths and weaknesses of both approaches for the F-formation task. They found that while the method of Cristani et al. [7] was more stable using head orientation information in the presence of noise, the method of Hung and Kröse [6] performed better when only position (and not orientation) information was available. Setti et al. [41] also proposed to handle the physical effect that different cardinalities of the F-formations sizes would have on the most plausible physical spatial layout of each member of the group. By taking this into account using separate accumulation spaces for each size, they were able to improve over the original Hough voting strategy proposed in [7]. A similar density-based approach has also been proposed by Gan et al. [3] where the final purpose of the task was to dynamically select camera angles for automated event recording. Tran et al. have subsequently analyzed temporal patterns of activities [10].

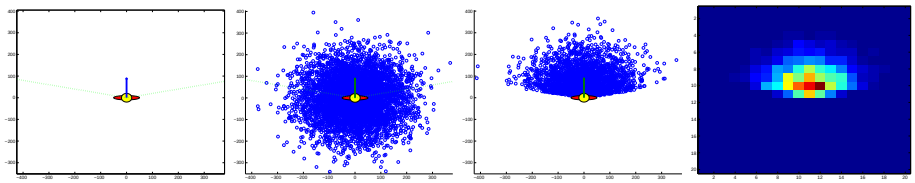
3 Our approach

Given a dataset of frames with positions of the persons and head/body orientations, the pipeline of the algorithm can be summarized in the following steps:

1. For each person $p_i \in P$ in a frame/scene, generate a frustum f_i based on his position and orientation as modeled by a 2-dimensional histogram (see Sec. 3.1)
2. Compute a pairwise affinity matrix for each $p_i \in P$ (see Sec. 3.2)
3. Extract F-formation (clusters) using evolutionary stable strategy-clusters (see Sec.4)

3.1 Frustum of attention modeling

Our frustum of social attention is inspired by Kendon's definition of a transactional segment. This takes into account both the field of view of the person and also the locus of attention of all other senses for a given body orientation. Since it is typically easier to obtain head pose rather than body or gaze orientation in crowded environments (due to occlusions), the head pose provides an approximation of the direction of the social attention frustum. It is characterized by a direction θ (which is the person's head orientation), an aperture α (we used $\alpha = 160^\circ$ which was reported by Ba and Odobez [38], who used the same measure for approximating the range of possible eye gaze directions given a specific head pose) and a length l. These three elements determine the socio-attentional frustum of a person. Given the parameters (θ, α and l) the frustum is modeled as a 2-dimensional (x and y position in the ground plane) Gaussian distribution in which each of the dimensions are generated independently. The parameter l corresponds to the variance of the Gaussian distribution centered around the location of a person. Therefore, a denser sampling is possible at locations closer to the person and decrease in density further away (after $3 * \sigma$ the number of sample are close to zero). The frustum is generated by drawing samples from the above Gaussian kernel



(a) (b) (c) (d)

Fig. 3. In figure is shown the process of generating the frustum: a) given the i-th person position and orientation a cone of aperture $\alpha = 160°$ is over imposed b) the 2D Gaussian set of samples are generated c) only the biologically feasible samples are kept d) binning of the space on a 20×20 grid to get the final histogram representation h_i.

and by keeping only those that fall within the cone (see Fig. 3). Given a person located at $p(x, y)$, with head orientation θ, a sample $s(x, y)$ is inside the frustum if

$$acos\left(\frac{s \cdot f_L}{||s|| * l}\right) \leq \frac{\alpha}{2} \tag{1}$$

where $f_L = \{\cos(\theta) * l, \sin(\theta) * l\}$ is the line of symmetry of the frustrum, a vector of length l and angle θ. This sampling process is iterated until the desired number of samples that falls in the cone is reached. The region that these samples represent intuitively models the transactional segment of a person. Each person in a scene is thus modeled using his frustum represented as 2-dimensional histogram h_i of size $N_c \times N_r$ normalized by the number of samples (s), where N_c and N_r span over the area of the scene captured by the camera.Experimentally, changing the value of the granularity such that $N_c \times N_r = 400, 2500, 10000$ did not change the overall performance (on the benchmarks we tried). Therefore, we keep the granularity fixed at 400 bins.

3.2 Quantifying pairwise interactions

Two persons are more likely to be interacting if their social attention frustums overlap. By quantifying the pairwise interaction as a distance between distributions, we are able to encode the uncertainty about the true transactional segment of the person given their head pose. Since we are dealing with histograms that represent discrete probability distributions, it is natural to consider information-theoretic measures to model the distance between them.

Given a pair of discrete probability distributions $P = \{p_1, \ldots, p_n\}$ and $Q = \{q_1, \ldots, q_n\}$, the first natural choice to measure their distance is given by the well-known Kullback-Leibler (KL) divergence, which is defined as:

$$D(P||Q) = \sum_{i=1}^{n} \log p_i \frac{p_i}{q_i} \tag{2}$$

The KL-divergence is known to be asymmetric. A symmetric version of the KL-divergence measure is the Jensen-Shannon (JS) divergence, which is defined as:

$$J(P, Q) = \frac{D(P||M) + D(Q||M)}{2} \tag{3}$$

where $M = \frac{1}{2}(P + Q)$ is the mid-point between P and Q. Hence, given two persons i and j in a scene and their vectorized histograms h_i and h_j, the distance between i and j can be calculated either as $D(h_i||h_j)$ or as $JS(h_i, h_j)$.

To obtain a measure of affinity, rather than distance, between each pair of histograms we used the classical Gaussian kernel:

$$\gamma(i,j) = \exp\left\{-\frac{d(h_i,h_j)}{\sigma}\right\} \tag{4}$$

where the funtion "d" refers to either the KL- or the JS-divergence. The parameter σ in Eq. 4 allows intrinsic properties of the scene (e.g., how far people usually stand from each other when they are in an F-formation) to be taken into account. Once we calculate this measure, it becomes possible to find groups of persons that are interacting by exploiting a grouping game, as described in the next section.

4 Grouping as a non-cooperative game

In this work we cast the approach proposed in [19,20] in the problem of detecting F-formations in terms of a non-cooperative *clustering game*. We choose this clustering algorithm for a series of nice properties:

- The distance function is not required to be symmetric, e.g. the Kullback-Leibler divergence.
- An a-priori number of clusters, like k-means, is not needed to be set since the algorithm let the cluster to emerge by data similarities. This represents a necessary condition since the number of groups in a scene is unknown.
- It search for maximal clique in a weighted graph which is an accepted definition of F-formation in the computer science community [6].
- Game-theory domain provides us the theoretical foundation to integrates multiple payoff matrices, thing of valuable importance when dealing with different temporal instants (see Sec.4.1).

Given a set of elements $O = \{1 \dots n\}$ and an $n \times n$ (possibly asymmetric) affinity matrix $A = (a_{ij})$ which quantifies the pairwise similarities between the objects in O, we envisage a situation whereby two players play a game which consists of simultaneously selecting an element from O. After showing their choice the players get a reward which is proportional to the similarity of the chosen elements. In game-theoretic jargon the elements of set O are the "pure strategies" available to both players and the affinity matrix A represents the "payoff" function (specifically, a_{ij} represents the payoff received by an individual playing strategy i against an opponent playing strategy j). In our application, the objects to to be grouped (namely, the pure strategies of this grouping game) correspond to the persons detected in a scene, the payoff function being the similarity measure between subjects as described in the previous sections.

A central notion in game theory is that of a *mixed strategy*, which is simply a probability distribution $\boldsymbol{x} = (x_1, \dots, x_n)^T$ over the set of pure strategies O. Mixed strategies clearly belong to the $(n-1)$-dimensional standard simplex:

$$\Delta = \left\{ \boldsymbol{x} \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1 \text{ and } x_i \geq 0,\ i = 1, \dots, n \right\}. \tag{5}$$

Given a mixed strategy $\boldsymbol{x} \in \Delta$, we define its *support* as $\sigma(\boldsymbol{x}) = \{i \in O : x_i > 0\}$.

The expected payoff received by an individual playing mixed strategy $\boldsymbol{y}$ against an opponent playing mixed strategy $\boldsymbol{x}$ is given by $\boldsymbol{y}^T A\boldsymbol{x}$. The set of *best replies* against a mixed strategy $\boldsymbol{x}$ is defined as $\beta(\boldsymbol{x}) = \{\boldsymbol{y} \in \Delta : \boldsymbol{y}^T A\boldsymbol{x} = \max_{\boldsymbol{z}} \boldsymbol{z}^T A\boldsymbol{x}\}$. Finally, a mixed strategy $\boldsymbol{x} \in \Delta$ is said to be a *Nash equilibrium* if it is a best reply to itself, namely if $\boldsymbol{x} \in \beta(\boldsymbol{x})$ or, in other words, if

$$\boldsymbol{x}^T A\boldsymbol{x} \geq \boldsymbol{y}^T A\boldsymbol{x} \tag{6}$$

for all $\boldsymbol{y} \in \Delta$. If inequality holds strictly, then $\boldsymbol{x}$ is said to be *strict* Nash equilibrium. Intuitively, at a Nash equilibrium no player has an incentive to unilaterally deviate from it. The clustering game is supposed to be played within an evolutionary setting wherein the two players, each of which is assumed to play a pre-assigned strategy, are repeatedly drawn at random from a large population. Here, given a mixed strategy $\boldsymbol{x} \in \Delta$, x_j ($j \in O$) is assumed to represent the proportion of players that is programmed to select pure strategy j. A dynamic evolutionary selection process will then make the population state $\boldsymbol{x}$ evolve according to a survival-of-the-fittest principle in such a way that, eventually, the better-than-average (pure) strategies will survive while the others will get extinct. Within this context, a mixed strategy $\boldsymbol{x} \in \Delta$ is said to be an *evolutionary stable strategy* (ESS) if it is a Nash equilibrium and if, for each best reply $\boldsymbol{y}$ to $\boldsymbol{x}$, we have $\boldsymbol{x}^T A\boldsymbol{y} > \boldsymbol{y}^T A\boldsymbol{y}$. Intuitively, ESS's are strategies such that any small deviation from them will lead to an inferior payoff (see [42] for an excellent introduction to evolutionary game theory).

In [19,20] a combinatorial characterization of ESS's is given which make them plausible candidates for the notion of a cluster (which they call ESS-cluster). The motivation behind this claim resides in the property that ESS-clusters do incorporate the two basic features which characterize a cluster, i.e.,

- *internal coherency*: elements belonging to the cluster should have high mutual similarities;
- *external incoherency*: the overall cluster internal coherency decreases by introducing external elements.

We refer to [19,20] for details. One of the distinguishing features of this approach is its generality as it allows one to deal in a unified framework with a variety of scenarios, including cases with asymmetric, negative, or high-order affinities. Note that, when the affinity matrix A is symmetric (that is, $A = A^T$) the notion of an ESS-cluster coincides with that of a dominant set [40], which amounts to finding a (local) maximizer of $\boldsymbol{x}^T A\boldsymbol{x}$ over the standard simplex Δ.

Algorithmically, to find an ESS-cluster one can use the classical *replicator dynamics* [42], a class of dynamical systems which mimic a Darwinian selection process over the set of pure strategies. The discrete-time version of these dynamics is given by the following update rule:

$$x_i(t+1) = x_i(t)\frac{(A\boldsymbol{x}(t))_i}{\boldsymbol{x}(t)^T A\boldsymbol{x}(t)} \tag{7}$$

for all $i \in O$. The process starts from a point $\boldsymbol{x}(0)$ usually close to the barycenter of the simplex Δ, and it is iterated until convergence (typically when distance between two

successive states is smaller than a given threshold). It is clear that the whole dynamical process is driven by the payoff function which, in our case, is defined precisely to favor the evolution of highly coherent objects. Accordingly, the support $\sigma(\boldsymbol{x})$ of the converged population state $\boldsymbol{x}$ does represent a cluster, the non-null components of which providing a measure of the degree of membership of its elements.

The support of an ESS corresponds to the indices of the elements in the same group. To extract all the ESS-clusters we implemented a simple peel-off strategy: when an ESS-cluster is computed the corresponding elements are removed from the original and the replicator dynamics is executed again on the remaining elements.

4.1 Integrating multiple frames in video sequences

When dealing with videos, the inter-frame smoothness between consecutive frames can be exploited to face cases of noisy data, such as wrong positions or head orientations. The idea is simply to consider a buffer of K frames: at time t, we will have knowledge of the frames at time $t-K+1,\ldots,t$, which can be used jointly for a more robust group estimation. This keeps the process of group modeling on-line (it can lie on top of the tracking algorithm), while permitting to prune out noise in an effective way. Assuming that the movement of the same person between frames is smooth, given a set of K consecutive frames, the problem is then to somehow integrate the corresponding affinity matrices to perform the grouping process.

From our game-theoretic perspective this problem can be seen in the context of multiple-payoff (or multi-criteria) games, a topic which has been the subject of intensive studies by game theorists since the late 1950's [43,44,45,46]. Under this setting, payoffs are no longer scalar quantities but take the form of vectors whose components represent different commodities. Clearly, the main difficulty which arises here is that the players' payoff spaces now can be given only a partial ordering. Although in "classical" game theory several solution concepts have been proposed during the years, the game theory community has typically given little attention to the evolutionary setting. Recently, a solution to this problem has been put forward by Somasundaram and Baras [21] who extended the notion of replicator dynamics and that of an ESS using the concept of Pareto-Nash equilibrium. Another recent attempt towards this direction, though more theoretical in nature, can be found in [47].

In the work reported in this paper, we follow the idea proposed in [21] which provides a principled solution to the problem of integrating multiple payoff functions. Using concepts from multi-criteria linear programming (MCLP) [48] they proposed a notion of Pareto reply and of Pareto-Nash equilibrium and showed the equivalence with "weighted sum scalarization", a classical technique from multi-objective optimization (see, e.g., [48]). Basically, this means that a Pareto-Nash equilibrium can be achieved by integrating the K affinity matrices as follows:

$$\hat{A} = \sum_{i=1}^{K} w_i A_i \tag{8}$$

where the w_i's ($i = 1 \ldots K$) represent appropriate non-negative trade-off weights associated to the different matrices, subject to the constraint $\sum_i w_i = 1$. Formulated in this

way, the problem of determining a Pareto-Nash equilibrium in a multi-payoff scenario is now reduced to the problem of determining the correct trade-off weights, and this in turn can be done by solving a multi-objective linear programming problem (MOLP). To this end, following [21], in our experiments we used the multi-objective simplex method (we refer the reader to chapter 7 of [48] and to the original paper [21] for details).

5 Experiments and results

We carried out experiments considering both the *single* (Sec. 5.3) and *multiple*-frame methods (Sec. 5.4) under ideal and noisy situation. In the former, F-formations are estimated on each single frame independently, while in the latter we perform integration over consecutive frames in order to smoothing noisy detection. Moreover we test the resilience of the method injecting increasing level of noise (Fig.5). Source code available at http://www.iit.it/en/datasets-and-code/code/gtcg.html

5.1 Datasets

The five datasets used (see Tab.1) are the currently publicly available benchmarks for detecting F-formations, where for each individual in a scene his x, y position and the head orientation are provided. Consecutive frames are available for two of them with a low frame rate. In three cases the annotation has been done via automatic tracking while other two were manually annotated by the respective authors as stated in Tab. 1.

PosterData [6]: it consists of 3 hours of aerial video of over 50 people during a scientific meeting involving poster presentations and a coffee break. 82 distinct image frames were selected based on maximizing differences between images, ambiguity in group membership and varying levels of crowdedness. 21 trained annotators were split into 8 trios who annotated 10-11 images for F-formations, leading to a subjective representation of the ground truth.

CocktailParty [5]: The CocktailParty dataset contains 16 minutes of video recordings of a cocktail party in a $30m^2$ lab environment involving 7 subjects. The party was recorded using four synchronized angled-view cameras (15Hz, $1024 \times 768px$, jpeg) installed in the corners of the room. The dataset is challenging for video analysis due to frequent and persistent occlusions, in a highly cluttered scene. Subject's positions and horizontal head orientations were logged using a particle filter-based body tracker with head pose estimation. Groups in one frame every 3 seconds were annotated manually by an trained expert, resulting in a total of 320 distinct frames for evaluation.

CoffeeBreak [7]: The dataset focuses on a coffee-break scenario of a social event, with max 14 individuals organized in groups of 2-3 people. People's positions were estimated by exploiting multi-object tracking on the heads, and head detection has been performed afterward, considering solely 4 possible orientations (Front, Back, Left, Right). The tracked positions were projected onto the ground plane. A trained expert annotated the videos indicating the groups present in the scenes (in combination with questionnaires

Dataset	#Sequences	#Frames × seq.	Consecutive Frames	Automated Tracking
CoffeeBreak	2	45,74	Y	Y
CocktailParty	1	320	Y	Y
GDet	5	132,115,79,17,60	N	Y
PosterData	82	1	N	N
Synth	10	10	N	N

Table 1. Datasets: multiple #Frame indicate diverse sequences, in these cases the final results are averaged over the sequences and normalized by the number of frames.

that the subjects filled in about the number of people they spoke with)on two different coffee breaks, for a total of 45 frames for *Seq1* and 75 frames for *Seq2*, acquired in a period of 3 seconds.

Synth [7]: A trained expert synthesised 10 different *situations*, with F-formation and singletons Each situation is repeated 10 times, with slightly varying position and orientation of the subjects. Here, noise in the position and orientations are absent.

GDet [7]: these videos consider a vending machines area where people take coffee and other drinks, and chat. In this case the head orientation considers solely 4 possible alternatives. Here the frame rate is very low, so that the multiple frame approach cannot be applied.

As comparative approaches, we consider the Hough-based approach of [7] in its renewed version of [18] (HFF), the hierarchical extension of the Hough-based approach of [41] (MULTI), and the dominant-set-based technique of [6](DS). Comparison with other baselines are not reported in Tab. 2 since are already carried out and overcomed in [18,7].

5.2 Evaluation metrics and parameter exploration

In terms of evaluation, as in [18], we consider a group as correctly estimated if at least $\lceil (T \cdot |G|) \rceil$ of their members are found by the grouping method were correctly detected by the algorithm, and if no more than $1 - \lceil (T \cdot |G|) \rceil$ false subjects are identified, where $|G|$ is the cardinality of the labeled group G, and $T = 2/3$. Based on this metrics, we produce *precision*, *recall* and *F measure* per frame; averaging these values over the frames gives the final scores.

Different combination of parameters are explored and validated on each dataset. In particular we examine the response of our approach when using the similarity functions (Eq. 2 and 3), by changing the value of $\sigma = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 0.9\}$ and the length of the frustum $l = \{20, 25, 30, 40, 50, 60, 80, 150\}$.

5.3 Single frame experiment

Tab. 2 shows the parameters used and the quantitative results obtained in the single-frame modality while in Fig.4 qualitative results of our group detector is shown com-

	CoffeeBreak (S1+S2)			PosterData			Gdet		
Method	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
HFF [18]	0,82	0,83	0,82	**0,93**	**0,96**	**0,94**	0,67	0,57	0,62
DS ([6], [18]*)	0,68	0,65	0,66	**0,93**	0,92	0,92	-	-	-
MULTISCALE [41]	0,82	0,77	0,80	-	-	-	-	-	-
Our KL	0,80	0,84	0,82	0,90	0,94	0,92	**0,76**	**0,75**	**0,75**
	σ=0.2 , l=40			σ=0.2 l=30			σ=0.5 l=80		
Our JS	**0,83**	**0,89**	**0,86**	0,92	**0,96**	**0,94**	**0,76**	**0,76**	**0,76**
	σ=0.2 , l=50			σ=0.3 , l=25			σ=0.5 l=80		

	Cocktail Party			Synth		
Method	Prec	Rec	F1	Prec	Rec	F1
HFF ([7], [41])	0,59	0,74	0,66	0,73	0,83	0,78
MULTISCALE [41]	0,69	0,74	0,71	0,86	0,94	0,90
Our KL	**0,85**	**0,81**	**0,83**	**1,00**	**1,00**	**1,00**
Our JS	**0,86**	**0,82**	**0,84**	**1,00**	**1,00**	**1,00**
	σ=0.5 , l=60			σ=0.1 , l=30		

Table 2. Results on single frame: only the best results are shown while the parameters are discussed in the paper (σ in Eq.4 and l in Eq.1). The comparative methods are: HFF [7], DS [18], MULTISCALE [41], JS or KL is our method using respectively the Jensen-Shannon (Eq.4) and the Kullback-Leibler (Eq.3) divergence. Maximum value for standard deviation for precision is 0.74% and for recall is 0.75%. * Note that in [18] the parameters for the DS method were not fully optimised.

pared with other method. As done in the comparative approaches, we show here the performances obtained with the best parameter settings, using both the Kullback-Leibler divergence (KL) and the Jensen-Shannon (JS) and averaged over 5 runs to evaluate the stability. As shown, the only case where we do not outperform the state of the art is on the Poster Data, with a difference of 1% in the precision with respect to HHF[18] and DS [6], a difference which is close to the maximum estimated variance of our approach. In the other cases, the results are definitely superior, saturating for example the synthetic benchmark, and outperforming by over 10% the *F-measure* on the GDet and the CocktailParty. It is worth noting that the performances across the different runs of the algorithm have been quite stable, with a mean standard deviation of $\simeq 0.74\%$ for both the precision and recall values.

5.4 Multiple frame experiment

The results are reported in Fig.5. Compared with the single-frame approach, in a noiseless tracking situation (blue curve), this version gives comparable results. As shown, the temporal integration varies almost uniformly except a slight increase in the CoffeeBreak Seq.1. In the case of noise (green,red and cyan curves) the single frame (first point on the curves) provides a low F score and is completely dominated by the multiframe version, irrespective of the number of frames considered in the buffer. To emphasize this fact a noise analysis on the CoffeeBreak and Cocktailparty datasets has been done. In these sequences, to simulate cluttered situations or noisy detector, we

(a) Seq 1 (b) Seq 1 (c) Seq 2 (d) Seq 2

Fig. 4. Qualitative results on the CoffeeBreak dataset compared with the state of the art HFF [7]. In yellow the groundtruth, in green our method and in red HFF. As evident from (a,b,c,d) HFF often fails in detecting groups of more than two persons while our approach is more stable.

injected noise in the orientation of persons by randomly selecting the frames to corrupt and the number of people to consider. In particular, the added orientation noise (γ) was 0-mean Gaussian, with a standard deviation varing in $\{\frac{\pi}{8}, \frac{\pi}{4}, \frac{\pi}{2}, \frac{2}{3}\pi\}$. The amount of frames and persons affected by noise was set by selecting from these percentages: $F = \{0\%, 25\%, 50\%, 75\%\}$, where the percentages indicate both the number of frames to be corrupted (whose time indexes have been sampled uniformly without replacement from the entire sequence) and the number of people affected by the noise. For example, in a sequence with 100 frames and 8 persons, setting a noise of 25% means to have 25 random frames where 2 random individual per frame are affected by noise. Considering the following size of the window $K = \{1, 2, 3, 4, 5\}$ of frames, we explore our approach applying the temporal integration.The Jensen-Shannon divergence has been used to generate the similarity matrices because it produces the better results in the single-frame experiments, outperforming the KL divergence in both the datasets. To combine the different similarity matrices in a buffer of K-frames, we used the average of the possible weights produced by the algorithm (Sec.4.1) normalized by their sum.

5.5 Discussion

Having these experimental evidences we can provide an overall final analysis. The proposed approach is to be preferred over the others under a wide variety of different scenarios. The performance are incredibly stable under both noisy (real) and ideal (synthetic) set. For example we have highest performance in the CoffeeBreak even if it is a very noisy dataset in terms of head orientation since only 4 orientations are possible while the Synthetic is an ideal case in which we reach 100% in precision and recall. From the single frame experiments it is clear that the Jensen-Shanon measure produces the highest and more stable performance. This seems to suggest that, while modeling a pairwise social interaction, it is reasonable to assume that both the individuals want to maintain a connection with the same strength, implying a symmetric affinity. Moreover the comparison between the multi frame and the single frame with noise reveals the meaningfulness of considering consecutive instants of the same scene to strengthen noisy detections. Concluding the blocks that absolutely contributed the most in this

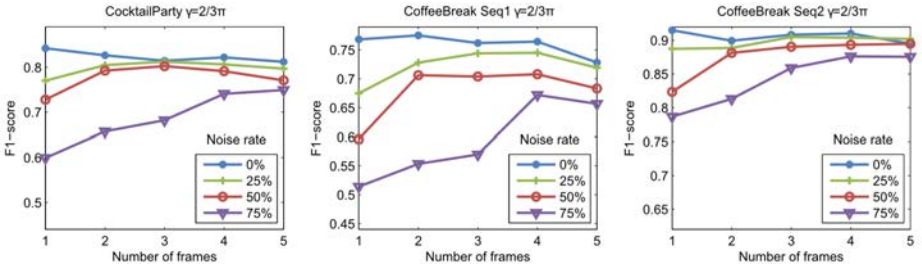


Fig. 5. Multiple-frame results: lines report the multiple-frame approach, with different level of noise: 0%, 25% , 50% and 75%. In this plot we show the worst case in which noise variance $\gamma = \frac{2}{3}\pi$. As visible, when noise is injected, the multiple-frame consistently outperforms the single-frame approach (first point of each curve). Mean value of the standard deviation for the precision is 1.61% and for recall is 1.73%.

work and that represents the main novelty, has been the biologically inspired model of the frustum, which capture far better the sociological interaction between individual with respect to the previous approaches, and the game-theoretic temporal integration which provides a principled way to efficiently prune noise by smoothing data across multiple frame.

6 Conclusions

In this paper we have proposed a new method for detecting conversational groups (F-Formations) that can be included in a typical surveillance pipeline or on top of a persons detector. The method has been designed to cope with very diverse realistic scenarios, dealing with both single/multi frame sequences, noisy tracking, missing detections, inaccurate face orientations and groups of any cardinality. This impacts several domains, like surveillance & security, behavior analysis, group detection, scene understanding and social signal processing. The approach improves upon existing methods by building a stochastic model of social attention from which the probability of an o-space existing between candidate pairs can be quantified using entropic measures.The resulting affinity matrix turns out to be more accurate than the ones used in the literature outperforming the actual state of the art. F-formations are extracted using a game-theoretic clustering approach which is able to efficiently find coherent groups in edge-weighted graphs. This game-theoretic perspective allowed us to integrate in a principle way the information coming from multiple consecutive frames, in an attempt to deal with noisy situations, like in a crowded scenario or due to inaccuracy of the detection algorithms. Our extensive experiments on single-frame showed improvements over other methods on five different datasets, while the integration with multiple frames allowed to augment the overall group detection accuracy, especially in the case of strong noise. Moreover encoding the frustum using an histogram makes the approach non-parametric and thus able to accommodate newer frustum models without changing the rest of the method. In the future, we plan to address the problem of modeling F-formations more deeply by considering points of instability when people leave or join groups and to integrate multiple cues (like gaze or body orientation) during the grouping process.

References

1. Groh, G., Lehmann, A., Reimers, J., Frieß, M.R., Schwarz, L.: Detecting social situations from interaction geometry. In: Social Computing (SocialCom), 2010 IEEE Second International Conference on, IEEE (2010) 1–8 1, 4
2. Li, R., Porfilio, P., Zickler, T.: Finding group interactions in social clutter. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2013) 1, 4
3. Gan, T., Wong, Y., Zhang, D., Kankanhalli, M.S.: Temporal encoded F-formation system for social interaction detection. In: Proceedings of the 21st ACM international conference on Multimedia. MM '13, New York, NY, USA, ACM (2013) 937–946 1, 5
4. Marin-Jimenez, M., Zisserman, A., Ferrari, V.: Here's looking at you, kid. detecting people looking at each other in videos. In: British Machine Vision Conference. (2011) 1, 2
5. Zen, G., Lepri, B., Ricci, E., Lanz, O.: Space speaks: towards socially and personality aware visual surveillance. In: 1st ACM international workshop on Multimodal pervasive video analysis. (2010) 37–42 1, 3, 10
6. Hung, H., Kröse, B.: Detecting F-formations as dominant sets. In: ICMI. (2011) 1, 2, 3, 5, 7, 10, 11, 12
7. Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Del Bue, A., Menegaz, G., Murino, V.: Social interaction discovery by statistical analysis of F-formations. In: Proc. of BMVC, BMVA Press (2011) 23.1–23.12 1, 2, 3, 4, 5, 10, 11, 12, 13
8. Lan, T., Wang, Y., Yang, W., Robinovitch, S.N., Mori, G.: Discriminative Latent Models for Recognizing Contextual Group Activities. IEEE Trans. Pattern Anal. Mach. Intell. **34** (2012) 1549–1562 1, 4
9. Yu, T., Lim, S., Patwardhan, K.A., Krahnstoever, N.: Monitoring, Recognizing and Discovering Social Networks. In: CVPR. (2009) 1, 4
10. Tran, K., Gala, A., Kakadiaris, I., Shah, S.: Activity Analysis in Crowded Environments Using Social Cues for Group Discovery and Human Interaction Modeling. Pattern Recognition Letters (2013) 1, 4, 5
11. Kendon, A.: Conducting Interaction: Patterns of Behavior in Focused Encounters (Studies in Interactional Sociolinguistics). Cambridge University Press (1990) 1, 4
12. Hüttenrauch, H., Eklundh, K.S., Green, A., Topp, E.A.: Investigating spatial relationships in human-robot interaction. In: Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on, IEEE (2006) 5052–5059 1
13. Hall, E.T.: The Hidden Dimension. Anchor (1990) 2
14. Goffman, E.: Behavior in Public Places: Notes on the Social Organization of Gatherings. Free Press (1966) 2
15. Ciolek, T.M., Kendon, A.: Environment and the Spatial Arrangement of Conversational Encounters. Sociological Inquiry **50** (1980) 237–271 2
16. Chen, C., Odobez, J.: We are not contortionists: coupled adaptive learning for head and body orientation estimation in surveillance video. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 1544–1551 2
17. Jain, V., Crowley, J.L.: Head pose estimation using multi-scale gaussian derivatives. In: Image Analysis. Springer (2013) 319–328 2
18. Setti, F., Hung, H., Cristani, M.: Group Detection in Still Images by F-formation Modeling: a Comparative Study. In: WIAMIS. (2013) 2, 5, 11, 12
19. Torsello, A., Rota Bulò, S., Pelillo, M.: Grouping with asymmetric affinities: A game-theoretic perspective. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). Volume 1. (2006) 292–299 3, 7, 8
20. Rota Bulò, S., Pelillo, M.: A game-theoretic approach to hypergraph clustering. IEEE Trans. Pattern Anal. Machine Intell. **35** (2013) 1312–1327 3, 7, 8

21. Somasundaram, K., Baras, J.S.: Achieving symmetric Pareto Nash equilibria using biased replicator dynamics. In: 48th IEEE Conf. Decision Control. (2009) 7000–7005 3, 9, 10
22. Pellegrini, S., Ess, A., Gool, L.V.: Improving data association by joint modeling of pedestrian trajectories and groupings. In: European Conference on Computer Vision (ECCV). (2010) 452–465 3
23. Yamaguchi, K., Berg, A., Ortiz, L., Berg, T.: Who are you with and where are you going? In: IEEE Conference on Computer Vision and Patter Recognition (CVPR). (2011) 3
24. Ge, W., Collins, R.T., Ruback, R.B.: Vision-based analysis of small groups in pedestrian crowds. IEEE Trans. on Pattern Analysis and Machine Intelligence **34** (2012) 1003–1016 3, 4
25. Qin, Z., Shelton, C.R.: Improving multi-target tracking via social grouping. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2012) 3
26. Chang, M., Krahnstoever, N., Ge, W.: Probabilistic group-level motion analysis and scenario recognition. In: IEEE ICCV. (2011) 3
27. Leal-Taixé, L., Pons-Moll, G., Rosenhahn, B.: Everybody needs somebody: modeling social and grouping behavior on a linear programming multiple people tracker. IEEE International Conference on Computer Vision Workshops (ICCVW). 1st Workshop on Modeling, Simulation and Visual Analysis of Large Crowds (2011) 3
28. Mckenna, S.J., Jabri, S., Duric, Z., Wechsler, H., Rosenfeld, A.: Tracking groups of people. Computer Vision and Image Understanding (2000) 3
29. Cupillard, F., Brémond, F., Thonnat, M., Antipolis, I.S., Group, O.: Tracking groups of people for video surveillance. In: University of Kingston (London). (2001) 3
30. Marques, J.S., Jorge, P.M., Abrantes, A.J., Lemos, J.M.: Tracking groups of pedestrians in video sequences. In: IEEE Conference on Computer Vision and Patter Recognition Workshops (CVPR workshops). Volume 9. (2003) 101–101 3
31. Pellegrini, S., Ess, A., Schindler, K., Gool, L.J.V.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: ICCV'09. (2009) 261–268 4
32. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Physical review E **51** (1995) 4282 4
33. Smith, K., Ba, S.O., Odobez, J.M., Gatica-Perez, D.: Tracking the Visual Focus of Attention for a Varying Number of Wandering People. IEEE Transactions on Pattern Analysis and Machine Intelligence **30** (2008) 1212–1229 4
34. Adams, R.B.: The science of social vision. Volume 7. Oxford University Press (2011) 4
35. Kendon, A.: Some functions of gaze-direction in social interaction. Acta Psychol (Amst) **26** (1967) 22–63 4
36. Jovanovic, N., op den Akker, R.: Towards automatic addressee identification in multi-party dialogues. In: Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue, Pennsylvania, USA, Association for Computational Linguistics (2004) 89–92 Imported from HMI. 4
37. Duncan, S.: Some signals and rules for taking speaking turns in conversations. Journal of Personality and Social Psychology **23** (1972) 283–292 4
38. Ba, S.O., Odobez, J.: Multiperson visual focus of attention from head pose and meeting contextual cues. Pattern Analysis and Machine Intelligence, IEEE Transactions on **33** (2011) 101–116 4, 5
39. Subramanian, R., Staiano, J., Kalimeri, K., Sebe, N., Pianesi, F.: Putting the pieces together: Multimodal analysis of social attention in meetings. In: Proceedings of the International Conference on Multimedia. MM '10, New York, NY, USA, ACM (2010) 659–662 4
40. Pavan, M., Pelillo, M.: Dominant sets and pairwise clustering. IEEE Trans. Pattern Anal. Machine Intell. **29** (2007) 167–172 5, 8

41. Setti, F., Lanz, O., Ferrario, R., Murino, V., Cristani, M.: Multi-Scale F-Formation Discovery for Group Detection. In: International Conference on Image Processing (ICIP). (2013) 5, 11, 12
42. Weibull, J.W.: Evolutionary Game Theory. MIT Press, Cambridge, MA (2005) 8
43. Blackwell, D.: An analog of the minimax theorem for vector payoffs. Pacific J. Math. **6** (1956) 1–8 9
44. Shapley, L.S.: Equilibrium points in games with vector payoffs. Naval Res. Logistics Quarterly **6** (1959) 57–61 9
45. Contini, B.M.: A decision model under uncertainty with multiple objectives. In Mensch, A., ed.: Theory of Games: Techniques and Applications. New York (1966) 9
46. Zeleny, M.: Games with multiple payoffs. Int. J. Game Theory **4** (1975) 179–191 9
47. Kawamura, T., Kanazawa, T., Ushio, T.: Evolutionarily and neutrally stable strategies in multicriteria games. IEICE Trans. Fundam. Electr. Commun. Comp. Sci. **E96-A** (2013) 814–820 9
48. Ehrgott, M.: Multicriteria Optimization. Springer, Berlin (2005) 2nd edition. 9, 10