Detecting Conversational Groups in Images and Sequences: a Robust Game-Theoretic Approach

Sebastiano Vascon[a,*], Eyasu Z. Mequanint[b], Marco Cristani[a,c], Hayley Hung[d], Marcello Pelillo[b], Vittorio Murino[a,c]

[a] *Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia, Genova, Italy*
[b] *Dept. of Environmental Sciences, Informatics and Statistics, University Ca' Foscari of Venice, Italy*
[c] *Dept. of Computer Science, University of Verona, Italy*
[d] *Faculty of Electrical Engineering, Mathematics and Computer Science, Technical University of Delft, Netherlands*

Abstract

This is the accepted version of the article, the official published version is available here https://doi.org/10.1016/j.cviu.2015.09.012

Detecting groups is becoming of relevant interest as an important step for scene (and especially activity) understanding. Differently from what is commonly assumed in the computer vision community, different types of groups do exist, and among these, standing conversational groups (a.k.a. F-formations) play an important role. An F-formation is a common type of people aggregation occurring when two or more persons sustain a social interaction, such as a chat at a cocktail party. Indeed, detecting and subsequently classifying such an interaction in images or videos is of considerable importance in many applicative contexts, like surveillance, social signal processing, social robotics or activity classification, to name a few. This paper presents a principled method to approach to this problem grounded upon the socio-psychological concept of an F-formation. More specifically, a game-theoretic framework is proposed, aimed at modeling the spatial structure characterizing F-formations. In other words, since F-formations are subject to geometrical configurations on how humans have to be mutually located and oriented, the proposed solution is able to account for these

*Corresponding author
Email address: sebastiano.vascon@iit.it (Sebastiano Vascon)

constraints while also statistically modeling the uncertainty associated with the position and orientation of the engaged persons. Moreover, taking advantage of video data, it is also able to integrate temporal information over multiple frames utilizing the recent notions from multi-payoff evolutionary game theory. The experiments have been performed on several benchmark datasets, consistently showing the superiority of the proposed approach over the state of the art, and its robustness under severe noise conditions.
Keywords: group detection, f-formation detection, conversational groups, game-theory, scene understanding

1. Introduction

The visual analysis of groups is becoming more and more widespread in computer vision, after decades of research on the automated modeling of individuals (which still remains an open problem), the goal has moved from encoding simple actions performed by a single subject to capturing dyads or clusters of social interactions [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. This is of extreme importance in many fields and applications, also addressing social and life sciences [11, 12]. This seems to be a necessary step, since humans are essentially a social species, as demonstrated by the fact that in everyday life people continuously interact with each other to achieve goals or simply to exchange states of mind. In this paper, we exploit a recent taxonomy presented in [13], which indicates that many types of groups can be defined. In particular, we target standing conversational groups, also known as *F-formations* [14], 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, such as video surveillance [7], social signal processing [2, 6, 4, 1], multimedia [3], social robotics [15], and activity recognition [16], 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 more socio-psychologically principled ways of designing methods for automati-

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

Figure 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.

cally analyzing human behavior. For example, Hall [17] proposed that relationships and levels of interactions could be inferred by considering different physical distances.

Goffman [18] 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 busy street, briefly greeting a colleague while passing them in the corridor, or indicating to let someone pass when boarding a train. This taxonomy has been exploited recently in [13] for addressing F-formations.

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" [19, p. 243]. Some examples of F-formations in real-world situations 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.

Three social spaces emerge from an F-formation: 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

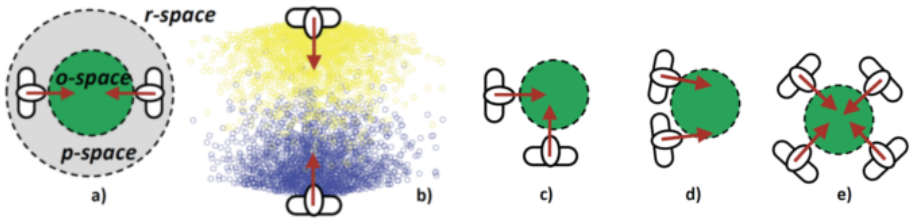


Figure 2: F-formations; a) components of an F-formation: o-space, p-space, r-space; in this case, a face-to-face F-formation is sketched; b) modeling the frustum of attention by particles: in the intersection stays the o-space; c) L-shape F-formation; d) side-by-side F-formation; e) circular F-formation.

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 outward 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 occlusions. These cues are easily obtainable nowadays, even if they are not estimated very accurately, and many approaches are aimed at extracting such information from raw images/videosequences [20, 21, 4]. Among the few approaches of F-formation detection, a recent experimental work of Setti et al. [22] shows that substantial improvement in the performance of F-formation detection algorithms can be achieved by combining a probabilistic approach (as [7]) and graph-based clustering methods [6]. Motivated by these studies, we develop a new sociologically-based approach which combines in a natural way the modeling of the uncertainty in the position and orientation of the subjects and a game-theoretic clustering approach , allowing one to extract coherent groups in edge-weighted graphs, digraphs and hyper-graphs [23, 24]. The game-theoretic setting provides a conceptual framework which allows us to integrate temporal information in a principled way, in an attempt to reliably extract groups in video sequences under severe noisy conditions. This is done by using a recent approach to integrate multiple payoff functions in an evolutionary

game-theoretic setting [25].

This work represents a substantial contribution to group detection in real scenarios. To date in computer vision, 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 [26, 27, 28, 29, 30, 31, 32, 33, 34]. In our case, F-formation are manifested primarily when people are still, so that a finer yet robust analysis is required. Our approach considers in fact the detection of groups in both still images and videos.

To test the effectiveness of the proposed approach, we performed extensive experiments over five different datasets, each one representing 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], the Cocktail Party [5] dataset and two new dataset, one proposed by Choi et al in [35] and FriendsMeet2 that we propose in this work. We also carried out systematic noise resilience experiments to fully investigate the stability and robustness of our method. The results consistently show the superior or comparable performances 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 been considerable prior work on estimating group activities by modeling behavior at the individual as well as group level [8, 9, 10, 36]. 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' [14]. These prior definitions of a group of people assume that they should be 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 [37] by the social force model [38], 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 [28, 36] or when people are conversing in a lecture hall [2]. Such an interaction requires a focusing of the senses, compared to the other group behaviors which can rely more on peripheral and unfocused sensing [18]: an interesting taxonomy of diverse kinds of social groups in relation to the kind of acquaintanceship of their members, and especially suited to computational frameworks, can be found in [13]. The automated analysis of different forms of unfocused and unfocused encounters was investigated extensively by Choi et al. [35] who created a data set of 7 different categories of group types, relating categories such as queues, sitting in a row, standing in a row, standing facing each other, and others. 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 behavior goes beyond just a cooperation between people to behave in a socially acceptable manner (e.g. by staying in line in a queue) and really indicates someone's willingness to be associated with someone else, and to actively exchange ideas and build social relations with a person.

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*[19]. 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 [39]. For visually led tasks such as looking at adverts [39], 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 [40]. More specifically, head pose is actually equally if not more perceptually salient as a cue for gaze direction in humans [40, 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 [41]. Jovanovic and Op den Akker also found that addressees could be identified using gazing cues [42], while Duncan found that speakers attracted the gaze of listeners [43] during conversations. Finally, Ba and Odobez [44] 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.

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 [45] 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. [22] 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. [46] also proposed to handle the physical effect that different cardinalities of the F-formations

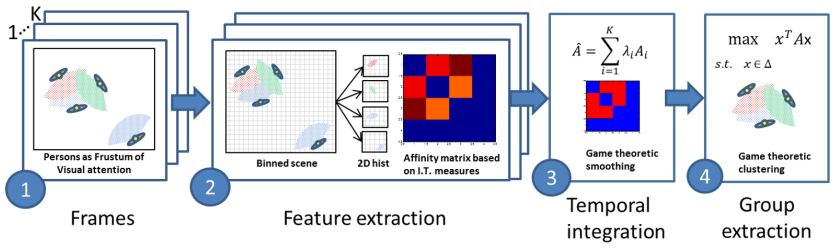


Figure 3: The pipeline of the algorithm

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]. Choi et al. [35] have modelled different forms of group behaviour discriminatively by projecting the body positions into 3d space and similar to our model, finding overlaps in a sampled density space. However, their approach was trying to distinguish differing group types and is not dedicated to conversational groups. Finally, Setti et al. [13] presented a graph-cut based minimization for detecting F-formations using proxemic data, that even if it shows strong performances, does not include temporal reasoning (it applies only to static images).

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 in world coordinates and modeled by a 2-dimensional histogram (see Sec. 3.1)
2. Compute a pairwise affinity matrix for each $p_i \in P$ (see Sec. 3.3)

3. In case a smoothing across multiple frame is required, compute the weights of each frame based on the theory of multipayoff games (see Sec.4.1)
4. Extract F-formation (clusters) using evolutionary stable strategy (ESS)-clusters (see Sec. 4)

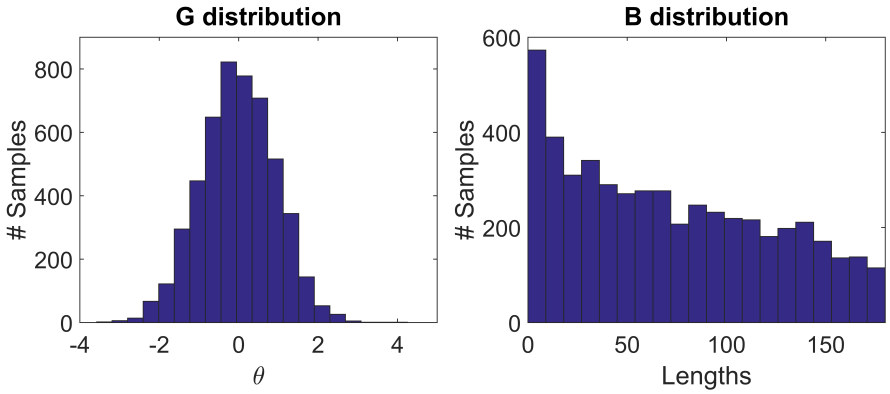
3.1. Frustum of attention modeling

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 the head pose rather than the body 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 γ (we used $\gamma = 160^{\circ}$ which was reported by Ba and Odobez [44], who used the same measure for approximating the range of possible eye gaze directions given a specific head pose) and a length l specified in cm or in $meters$ based on the data. These three elements determine the socio-attentional frustum of a person. In this work we propose a new frustum model based on sampling from two probability distributions. In our approach the sampling has a twofold impact in the whole pipeline:

- *application decoupling*: it decouples the entire algorithm from a specific model because using samples and histograms makes the entire approach non-parametric and thus able to easily accommodate forthcoming models;
- *data smoothing*: sampling methods, in general, smooth noisy data by looking for a statistical consensus. It is quite common in our scenario to deal with unreliable data since tracking, detections, head orientations, etc. are all noisy and prone to errors;

The new model differ from [47] in two fundamental aspects:

- *sampling method*: previously, samples were drawn from a 2D Gaussian distribution which was chopped based on the field of view (see Fig. 5a). Each sample was subsequently marked as valid or not and the drawning process stops until the



(a) (b)

Figure 4: a) The probability distribution over the orientations and b) the distance from the person.

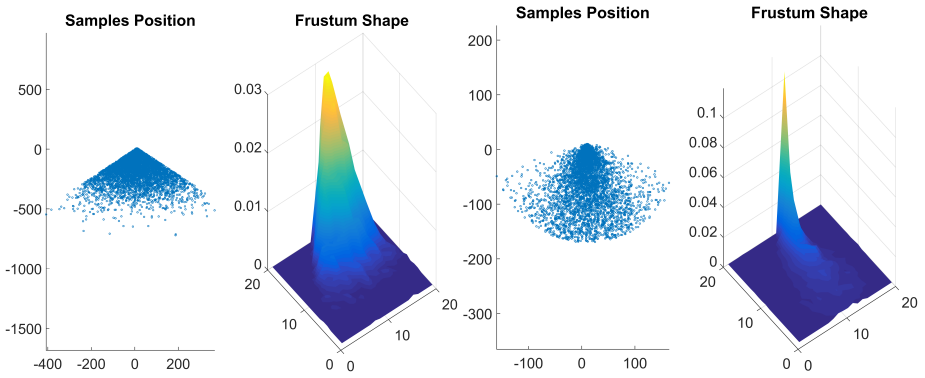
desired number of valid samples was reached. This approach is time consuming. The samples generated using our new method are all valid by default speeding up the entire process.

- *peripheral view*: the new method is more expressive since it is able to capture the peripheral field of view (see Fig. 5b) emboding the natural lateral decay of the human view instead of the sharp boundaries of the previous approach.

These two modifications are reflected into an higher performances and an overall speedup of the algorithm. More precisely, the proposed new frustum is based on a combination of two probability distributions, a Gaussian distribution G and a Beta distribution B.

The G *distribution* (see Fig4 a) is used to generate samples related to the aperture of the frustum so is centered in the head orientation θ of a person with a variance set such that the full width of the Gaussian distribution corresponds to the desired aperture of the frustum. In a Gaussian distribution the 99% of the samples are located in the range of $[-3\sigma, 3\sigma]$; this range will correspond to the full aperture of the frustum, so that setting the variance in a way that the aperture is fully covered becomes an easy task, $\sigma = \frac{1}{3} * \frac{\gamma}{2}$.

The B *distribution* (see Fig4 b) is used to generate samples that are dense in close



(a) (b)

Figure 5: a) The old frustum model proposed in [47]. b) the newer frustum based on the sampling from the two distribution.

proximity of the person while decades going far away, to achieve this shape we set the distribution parameters $\alpha = 0.8$ and $\beta = 1.1$ (see Fig5a). The values returned by the B distribution are bounded in $[0, 1]$ and need to be multiplied by the desired length of the frustum l. The samples obtained using these two distributions are in polar coordinates (an angle and a distance), to obtain samples in the 2D space is sufficient to apply a simple trigonometric rule to each of them. Given a pair of samples from G and B (G_i, B_i) and the position of a person (p_x, p_y) the 2D position of each sample is:

$$\begin{aligned} s_x &= p_x + \cos(G_i) * B_i * l \\ s_y &= p_y + \sin(G_i) * B_i * l \end{aligned} \tag{1}$$

Drawn independently n samples from both the distributions and applying the above equations we obtain a set of samples that falls in the human frustum of visual attention. With respect to the previous model there is no need to have a continuous sampling until the desired number of samples are reached, because all pairs of sample generated from the N and B distributions are already valid without the need of pruning the unbiological ones and making the approach faster.

Using the approach in [47] with the number of samples $n = 5000$ the time to generate the feasible samples is $\simeq 0.161s$ while with the new method is $\simeq 0.008s$, speeding

up the entire algorithm twenty time the previous approach. Each person in a scene is thus modeled using his/her frustum represented as 2-dimensional histogram h_i of size $N_c \times N_r$, normalized by the number of samples (n), where N_c and N_r span over the area of the scene captured by the camera.

3.2. Histogram binning

In order to decide the best binning of the 2D histogram we decided to carry out an extensive experimentation over all the publicly available datasets to see which size gives the better trade-off between required space to store the histograms and performances. We tested the performances using binning ranges from [5, 10, 15, 20, 30, 50, 100, 150, 200, 300, 400], obtaining the F1-score plotted in Fig.6.

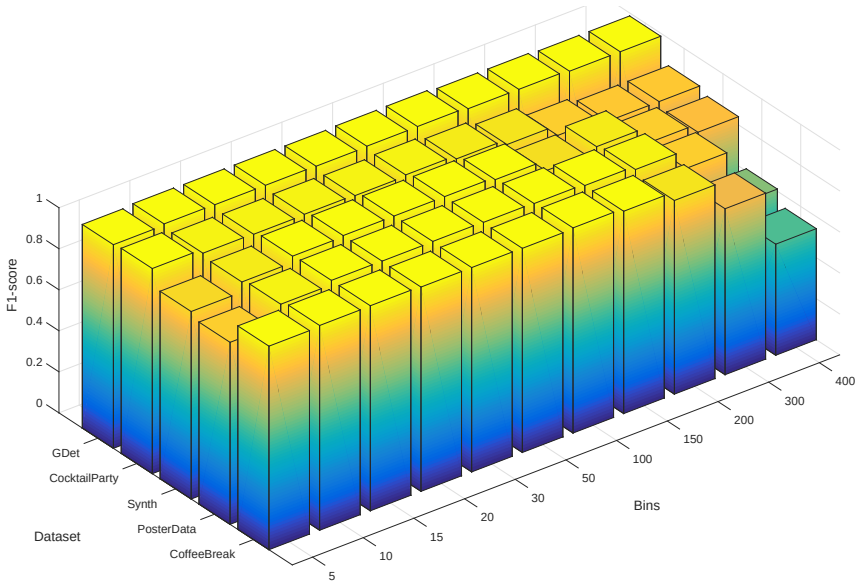


Figure 6: F1 score on each dataset by changing the number of bins in the histogram. As one can see the performances are quite stable if the binning varies from 10 to 100.

Summing the F1-score obtained on each dataset for a particular binning it is possible to rank the performance, obtaining the Table3.2: Here we ca see (Table3.2) that the best performances are obtained using $N_c = N_r = 50$ or 20 bins; since their performance difference is very low (0.002), we decided to use histograms of size 20x20 to keep the size as smaller as possible without losing the strengths of the method.

#Bins	Rank	#Bins	Rank	#Bins	Rank
50	4,974583	100	4,948771	200	4,757109
20	4,972243	10	4,918738	300	4,423196
15	4,96876	150	4,844176	400	3,843835
30	4,968515	5	4,786053		

Table 1: This table show the overall performances of the F1-score across different dataset using different binning. Rank are based on the sum of the F1-scores.

3.3. Quantifying pairwise interactions

Two persons are more likely to be interactants 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 persons 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 [48], 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:

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

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 [23] in the problem of detecting F-formations in terms of a non-cooperative *clustering game*. We choose this clustering algorithm for a series of desirable properties:

- The similarity function is not required to be a metric, so it is usable with the Kullback-Leibler.
- Setting an a-priori number of clusters, like in the k-means procedure, is not needed. This is useful, since the number of groups in a scene in unknown.
- Game-theory domain provides us the theoretical foundation to integrate multiple payoff matrices, which is of valuable importance when dealing with different temporal instants (see Sec.4.1).

Despite the above properties and for the sake of completeness, the performances of the game-theoretic clustering in this scenario has been compared with a more traditional method, the Spectral Clustering [49] algorithm, showing the superiority of the first method. The details of the experiment and the quantitative results has been reported respectively in Sec.5.3.3 and in Table 3 (see "R-GTCG SC" rows).

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 $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ over the set of pure strategies O. Mixed strategies clearly belong to the $(n-1)$-dimensional standard simplex:

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

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 [50] for an excellent introduction to evolutionary game theory).

In [23, 24] a combinatorial characterization of ESS's is given which makes 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 [23, 24] 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 [45], 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* [50], 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.

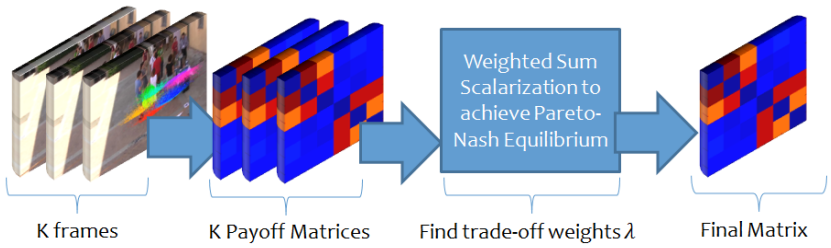


Figure 7: Pipeline for the multipayoff approaches.

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 [51, 52, 53, 54]. 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 set-

ting. Recently, a solution to this problem has been put forward by Somasundaram and Baras [25] 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 [55].

In the work reported in this paper, we follow the idea proposed in [25]. Using concepts from multi-criteria linear programming (MCLP) [56] 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., [56]). Basically, this means that a Pareto-Nash equilibrium can be achieved by integrating the K affinity matrices as follows:

$$\hat{A} \quad = \quad \sum_{i=1}^{K} \hat{w}_i A_i \tag{8}$$

where the $\hat{w}_i$'s ($i = 1 \ldots K$) represent appropriate non-negative trade-off weights associated to the different matrices, subject to the constraint $\sum_i \hat{w}_i = 1$. 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. To this end, following [25], in our experiments we used the multi-objective simplex method described in [56, Chapter 7] (see also [25] for details).

The algorithm described above provides as output a set of weight vectors which allows one to obtain the whole Pareto front of the original multi-objective problem. However, in practice, in a decision-making context like ours one has to obtain somehow a single solution but, in general, it is not clear how to do it since it might depend on subjective or other extra-criterion preferences on the decision maker's part. Below we provide some heuristics which are motivated by the following empirical observations.

1. If the matrices are all very similar to each other the weights generated are uniformly distributed meaning that the matrices are all equally important.
2. If a matrix is very different from the other ones (e.g. one noisy frame) it gets the highest weight.

At a first glance, the straightforward idea is to remove the matrix with the highest

weight, unfortunately this could lead to very bad results because the most diverse could be the matrix without corruption if two frames are taken into account. Given a set of weights $\mathbf{W} = \{\overline{w}_1, \ldots, \overline{w}_n\}$, in which each $\overline{w}_i$ is an K-dimensional vector representing the weight for each payoff matrix $A_{1\ldots K}$, we propose the following heuristics to select/generate the proper set $\hat{w}$:

4.1.1. Normalized sum of the weights

The set $\mathbf{W}$ can be seen as a $k \times n$ matrix in which each column is a feasible solution from the Pareto front. The final weights are computed summing the rows of the matrix $\mathbf{W}$ and normalizing the result by the sum, so that $\sum_{i=1}^{n} \hat{w}_i = 1$:

$$\begin{aligned} \hat{w_i} &= \sum_{j=1}^{n} \mathbf{W}_{j,i} \\ \hat{w_i} &= \frac{\hat{w_i}}{\sum_{i=1}^{K} w_i} \end{aligned} \tag{9}$$

this heuristic has been used in [47] to solve the problem of highly unbalanced weights that occur when few frames in the window are available (in particular in case of two frames). Few frames, in general, produce a reduced set of feasible solutions. Moreover, in the case of noise, the algorithm assigns an higher weight to the more dissimilar matrix (the correct or the corrupted ones), ending up into a set of unbalanced weights with the limit case of weights in $\{0,1\}$. For example, given two consecutive frames (A,B) with B the corrupted ones, a possible set of weights can be:

$$\mathbf{W} = \begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} \overline{w}_1 & \overline{w}_2 \end{array} \\ \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix} \end{array}$$

If the $\overline{w}_1$ is taken the uncorrupted matrix is chosen while taking $\overline{w}_2$ the opposite happens. Choosing $\overline{w}_2$ lead to a poor performances or wrong grouping since the corrupted frames is chosen. As one can see, making a decision in this context is very complicated because we have no prior information on whether the noise is and the weights are not informative enough. To avoid to make this choice, a statistical consensus is searched by meaning the possible solution. This heuristics, in the previous example, will assign

the same weights $\hat{w_A} = \hat{w_B} = 0.5$ to both the frames. Thus in the final matrix at least half of the correct data are kept from A and half of the corrupted data from B are left. In case that more frames are taken into account, the number of possible weights become larger and thus the true importance of each frame emerges naturally by meaning the weights.

4.1.2. Weighting of the solutions

The rationale of this approach is to weigh each solution $w \in \mathbf{W}$ based on the similarity with respect to the other; this leads the most similar one to be chosen as the most representative, while the others participate less in the final solution. A way of deciding the weights is using the characteristic vector provided by the Dominant Set [45] approach. A graph G=(V,E,ω) is created in which $V = W$, the set of weights, and the similarity between pairs of weights i, j is $\omega(w_i, w_j) = e^{-||w_i - w_j||_2}$. The first dominant set is extracted and the corresponding characteristic vector x is used to weigh each w. The final weight is:

$$\begin{aligned} \hat{w_i} &= \sum_{j=1}^{n} x_j * \mathbf{W}_{i,j} \\ \hat{w_i} &= \frac{\hat{w_i}}{\sum_{i=1}^{K} w_i} \end{aligned} \tag{10}$$

4.1.3. Maximal Entropy

The rationale is to select the set of weights which are close to an uniform distribution, in order to prune trivial solutions (the ones having only one matrix with full weight) and keeping the peculiar characteristics of different payoff matrices in which each matrix is represented.

$$\hat{w} = \arg\max_{w \in \mathbf{W}} \left(- \sum_i w_i \log_2 w_i \right) \tag{11}$$

4.1.4. Consensus of Clustering Ensemble

The rationale here is that, since each set of weights $w \in \mathbf{W}$ represents a feasible theoretical solution, we use each set separately to extract the groups of people from an image obtaining n different possible groupings. The final solution is generated

via "consensus" between the different groupings from the n solutions. Given a set of groupings, the consensus is found by an evidence accumulation matrix $E = m \times m$, similar to [57], in which m is the maximum number of persons in the scenes and $E_{i,j}$ counts the number of times that persons i and j are grouped together in n different solutions. The E matrix is then divided by the number of solutions n. The final grouping is then obtained through clustering over the matrix E using the Dominant Set approach.

5. Experiments and Results

We carried out experiments considering both the single- (Sec. 5.3) and multiple-frame methods (Sec. 5.4) in ideal and noisy situations. 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 filter out noisy detections. Moreover, the robustness of the method injecting increasing levels of noise (Fig.9) has been tested.

5.1. Datasets

The seven datasets used (see Tab. 2) are the current publicly available benchmarks for detecting F-formations, where for each individual in a scene its x, y position and the head orientation are provided. Consecutive frames are available for three of them with a low frame rate. In four cases the annotation has been done via automatic tracking while other two were manually annotated by the respective authors as stated in Tab. 2. The datasets that are used in this work are all in world coordinates, with a top-view camera setting, except for the Idiap Poster data, which used an orthogonal projections of a 3-D bounding box overlayed on people in the scene to determine the position of each person in the image plane.

FriendsMeet2 (FM2). It consists of an extensions of the FriendsMeet proposed in [58]; this version is composed by the 15 original real sequences, in which additionally the head orientation has been manually annotated in each sequence. This results in 10685 annotated frames, the biggest dataset for group detection available to date. The head orientation annotation has been done in the image plane by pointing to the head of the person and drawing a line in the direction where she/he is looking. Through the

available homography has been possible to convert the line from image plane to world coordinates obtaining the real head angle on the ground plane. The ground-truth for the groups is the same as in the original dataset.

Discovering Groups of People in Images (DGPI). This dataset has been recently proposed by Choi et al. [35] and is composed of 599 images of different real environments. Groups are divided into 7 categories based on the internal disposition of the persons: *queuing (Q), standing facing each other (SF), sitting facing-each-other (OF), sitting on the ground facing-each-other (GF), standing side by side (SS), sitting side by side (OS), sitting on the ground side by side (GS)*. For each image, the annotation of the persons position, head orientation, the groups and the corresponding types are provided. The dataset has been not specifically designed for the task of conversational groups detection and, on the basis of the hypothesis that facing is mandatory to converse, we use only the groups of facing people (SF, OF, and GF) as our ground-truth.

PosterData [6]. It consists of 3 hours of indoors video in the large atrium of an hotel building with over 50 people during a scientific meeting involving poster presentations and a coffee break. The cameras were mounted from above pointing downwards to record the scene. The 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 7 groups (3 persons each) 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. This scenario 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 given the highly cluttered scene. Subject's positions and horizontal head orientations were logged using a particle filter-based body tracker

with head pose estimation. Groups were annotated manually by a trained expert every 3 seconds, 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 positions were estimated by exploiting multi-object tracking of the heads, and head orientation 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 that the subjects filled in about the number of people they spoke with) on two different coffee-break events, for a total of 45 frames for *Seq1* and 75 frames for *Seq2*, acquired every 3 seconds.

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

GDet [7]. The scenario consists in a vending machines area where people take coffee and other drinks, and chat. In this case, head orientation considers solely 4 possible alternatives and, since the frame rate is very low, the multiple frame approach cannot be applied.

As comparative approaches, we consider the Hough-based approach of [7] in its renewed version of [22] (HFF), the hierarchical extension of the Hough-based approach of [46] (MULTI), the dominant set-based technique of [6](DS), and the approach of Choi et al [35]. Comparison with other baselines are not reported in Tab. 3 since they are already carried out and reported in [22, 7].

5.2. Evaluation metrics and parameter exploration

In terms of evaluation, as in [22], we consider a group as correctly estimated if at least $\lceil (T \cdot |G|) \rceil$ of their members are correctly detected by the algorithm, and if no more than $\lceil (1 - T) \cdot |G| \rceil$ false subjects are identified, where $|G|$ is the cardinality of

Dataset	#Sequences	#Frames	Consecutive Frames	Automated Tracking
FriendsMeet2	15	10685	Y	N
DGPI	1	599	N	Y
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 2: Datasets: multiple #Frame indicate diverse sequences, in these cases the final results are averaged over the sequences and normalized by the number of frames.

the labeled group G, and $T = 2/3$. The DGPI [35] dataset uses a different criterion to evaluate the performances: a group is correctly detected if at least half of the persons in a detected group matches a group in the ground-truth. In practice, it is the same criterion as above but with T parameter equal to 0.5. Based on this metrics, we compute *precision*, *recall*, and *F1-score* per frame; averaging these values over the frames gives the final scores.

Different combinations of parameters are explored and validated on each dataset. In particular, we examine the performance of our approach when using the similarity function in Eq.4 with the distance function in Eq.3 (as suggested by [47]), and by varying the value of σ in the range $\{0.1, 0.2, 0.4, 0.5, 0.7, 0.9\}$. To explore the effect of the length of the frustum we based our analysis on the studies conducted in [59, 60] in which a focused encounter between two persons may occur between 45 cm to 2 meters; correspondingly, the parameter l will range in the same interval.

5.3. Single-frame experiment

In this experiment, we apply our method on several publicly available datasets and one new dataset proposed in this paper. The section is subdivided into two parts, the

first in which we compare our method with consolidated public benchmarks, and the last one dedicated to the new datasets.

5.3.1. State-of-the-art datasets

Tab. 3 shows the parameters used and the quantitative results obtained in the single-frame modality, and in Fig.8 qualitative results of our group detector are shown in comparison with the HFF method [7]. As done in the comparative approaches, we show here the performances obtained with the best parameter settings using the Jensen-Shannon (JS) divergence, and averaged over 10 runs to evaluate the stability. As shown, the only cases where our approach does not outperform the state-of-the-art [6] is on the Poster Data, with a difference of 1% in the precision with respect to HHF, and in the recall of the CoffeeBreak with a difference of 1% , a difference which is still below the variance found in the experiments. In the other cases, the results are definitely superior, saturating for example the synthetic benchmark, and outperforming by over 10% the *F1-score* on the GDet and the CocktailParty. In the IRPM approach, results are considerable low, since the F/formation modeling was one of the first being formalized in computer vision from the sociological literature. In particular, it stated that a group is formed by people whose view frustum intersect, without accounting for occlusions among people. This generated many false positives that, in turns, cause the approach to lose many good groups (in other words, a group is estimated accounting the wrong people, which in turns may have been involved in other formations). It is worth noting that the performances across the different runs of the algorithm have been quite stable, with a maximum variance about 0.6% for both the precision and recall values.

5.3.2. New datasets

The results on the DGPI dataset are reported in Table 4. The features on this dataset are automatically annotated and for this reason this dataset is represents a very challenging test bed. The comparison with [35] has been made by taking as our ground-truth the union of the ground-truths of the facing persons, and we compared our results with the average on the related ground-truth obtained by [35]. The reported values

	CoffeeBreak (S1+S2)			**PosterData**			**Gdet**		
Method ⋆	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
IRPM [61],[22]	0.60	0,41	0,49	-	-	-	**-**	**-**	**-**
HFF [22]	0,82	0,83	0,82	**0,93**	**0,96**	**0,94**	0,67	0,57	0,62
DS ([6], [22])*	0,68	0,65	0,66	**0,93**	0,92	0,92	**-**	**-**	**-**
MULTISCALE [46]	0,82	0,77	0,80	-	-	-	**-**	**-**	**-**
GTCG [47] KL	0,80	0,84	0,82	0,90	0,94	0,92	**0,76**	0,75	0,75
GTCG [47] JS	0,83	**0,89**	0,86	0,92	**0,96**	**0,94**	**0,76**	**0,76**	**0,76**
R-GTCG SC	0,52	0,59	0,55	0,26	0,27	0,26	0,75	0,75	0,75
R-GTCG	**0,86**	0,88	**0,87**	0,92	**0,96**	**0,94**	**0,76**	**0,76**	**0,76**
	σ=0.2 , l=145			σ=0.25 , l=115			σ=0.7 l=180		

	Cocktail Party			**Synth**		
Method ⋆	Prec	Rec	F1	Prec	Rec	F1
IRPM [61],[22]	-	-	-	0,71	0,54	0,61
HFF ([7], [46])	0,59	0,74	0,66	0,73	0,83	0,78
MULTISCALE [46]	0,69	0,74	0,71	0,86	0,94	0,90
GTCG [47] KL	0,85	0,81	0,83	**1,00**	**1,00**	**1,00**
GTCG [47] JS	0,86	**0,82**	**0,84**	**1,00**	**1,00**	**1,00**
R-GTCG SC	0,77	0,72	0,74	0,40	0,90	0,56
	σ=0.6 , l=170			σ=0.1 , l=75		
R-GTCG	**0,87**	**0,82**	**0,84**	**1,00**	**1,00**	**1,00**

Table 3: 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: IRPM [61], HFF [7], DS [22], MULTISCALE [46], GTCG [47], "$R-GTCG$" our method and "R-GTCG SC" the results of our method using the Spectral Clustering technique instead of the game-theoretic-clustering. * Note that in [22] the parameters for the DS method were not fully optimised.⋆ In case of double citations, the first one refer to the original method while the latter refers to the more recent paper from which the results has been taken.

are calculated using the result from the full model in Table 2 of [35]. If all the seven ground-truth classes are considered as generic groups and used for evaluating our ap-

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

Figure 8: Qualitative results on the CoffeeBreak dataset compared with the state of the art HFF [7]. In yellow the ground-truth, 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.

	Prec	Rec	F1-score
Choi et al [35]	0.59	0.65	0.62
Our	0.54	0.57	0.56

Table 4: Results obtained on the DGPI dataset.

proach, we have very high figures with a precision reaching 0.99, a recall of 0.73, and an F1-score of 0.84. Unfortunately, in [35] only the average over the detections per type of groups is reported (precision 0.50, recall 0.44, and F1-score 0.47), so this comparison is not completely fair but it gives an indication of the goodness of our approach also in these conditions.

The results on the FM2 dataset are reported in Table 5, the parameters used in this dataset are the same for all the sequences, since there is no change of the viewpoint, $l = 1.5$ and $\sigma = 0.5$. As one can note, the performance is very low in three sequences (Seq 11, 12, and 13): this is motivated by the fact that in these sequences there are no conversational groups but persons that are queuing or walking producing a lot of false positives. If we keep only the sequences in which persons are facing and interacting (Seq1 to 10, Seq14 and Seq15), we obtain very good performances with a mean precision of 0.886 , recall equal to 0.862 , and F1-score equal to 0.873.

	Seq1	**Seq2**	**Seq3**	**Seq4**	**Seq5**
Prec	0.88441	0.92537	0.74896	0.85863	0.90859
Rec	0.82137	0.93419	0.74896	0.7578	0.87165
F1	0.85172	0.92976	0.74896	0.80507	0.88974
	Seq6	**Seq7**	**Seq8**	**Seq9**	**Seq10**
Prec	0.9289	0.91389	0.68278	1	0.97187
Rec	0.9289	0.91854	0.65559	0.99804	0.93095
F1	0.9289	0.91621	0.66891	0.99902	0.95097
	Seq11	**Seq12**	**Seq13**	**Seq14**	**Seq15**
Prec	0.41232	0.11231	0.12818	0.86818	0.93551
Rec	0.30806	0.13478	0.1409	0.86818	0.91408
F1	0.35264	0.12252	0.13424	0.86818	0.92467

Table 5: Results on the different sequences of the FriendsMeet2.

5.3.3. ESS-clustering vs. Spectral Clustering

The rationale of this experiment is to compare the overall performances of our pipeline when changing the clustering method in the last step. To this end a more traditional clustering technique, the spectral clustering [49] algorithm [1], has been chosen. In this experiment the last step (the fourth) of the pipeline has been changed, substituting the Game-Theoretic-clustering method with the Spectral Clustering and keeping the remaining steps fixed. The single-frame modality on all the state-of-the-art datasets has been explored since the interest is to prove the validity of our clustering choice rather than other steps of the algorithm. The evaluation criteria are exactly the same as the other experiments and the quantitative results have been reported in Table 3 (see "R-GTCG SC" rows) showing the superiority of the game-theoretic approach.

[1] Note that since the number of groups in a scene (number of clusters) are not known a-priori, the *spectral gap* heuristic [49] has been used to find the proper subdivision of the data

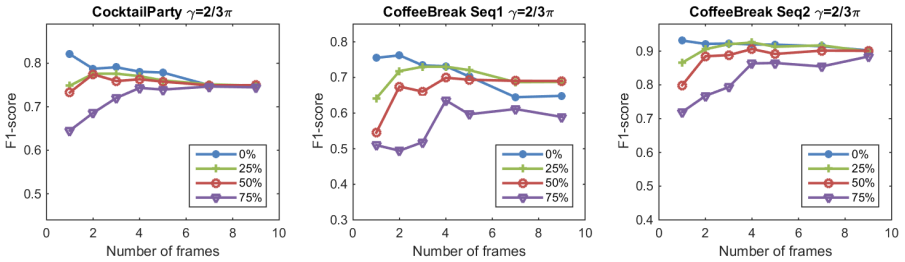
5.4. Multiple-frame experiment

In this experiment we carried out the analysis in case of noisy unreliable data, faced using the multi-payoff game theory proposed in Sec. 4.1. This analysis involves the injection of noise in the data, and in particular in the head orientation rather than in the person position; this because the orientation is the most problematic feature to be automatically extracted and so the more noisy. Given a window of K frames, the noise is added to the head orientation on an increasing number frames and persons in the scene, and is expressed in percentage. In particular, the amount of frames and persons affected by noise was set by selecting from these values: $F = \{0\%, 25\%, 50\%, 75\%\}$, where the percentages indicate both the number of corrupted frames (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 in which the head orientation of 2 (random) individuals is altered by noise. Considering the following size of the window $K = \{1, 2, 3, 4, 5, 7, 9\}$ of frames, we explore our approach applying the temporal integration. The JS divergence has been used to generate the similarity matrices because it produces the better results [47] in the single-frame experiments, outperforming the KL divergence in both the datasets.

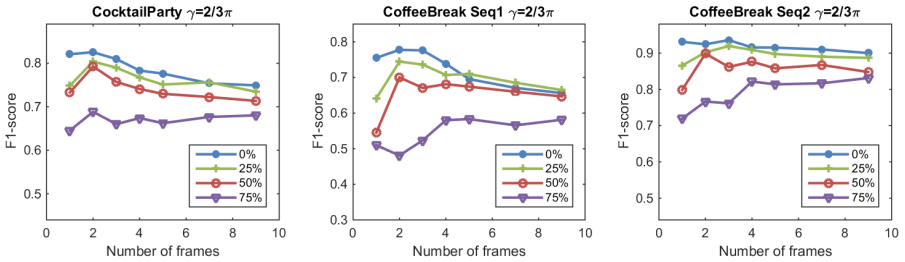
5.4.1. Heuristic performance analysis

In Fig.9, we analyze the performance of the heuristics proposed in Sec. 4.1 under the highest level of noise in the head orientation $\gamma = \frac{2}{3}\pi$, and on an increasing number of corrupted frames and persons. From the Fig. 9, we can drawn the following conclusions for each heuristic:

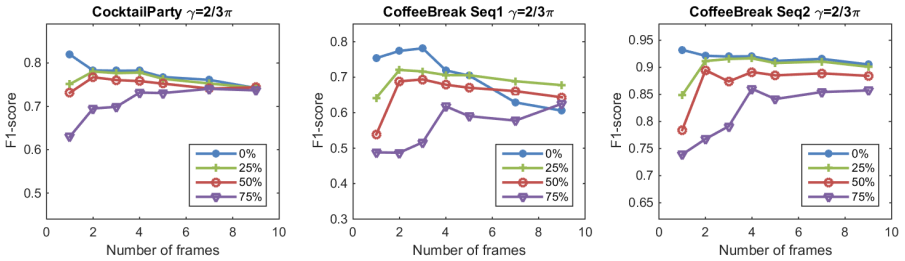
Fig. 9a . The normalized sum of weight is the best performing heuristic over the other methods, in particular when a small number of frames is considered. This is because when few frames (2 or 3) are taken into account, the number of possible solutions is more or less in the same order, if one keeps at random one of these solution it may choose exactly the corrupted frame so averaging over the possible weights makes an uniform distribution, eventually avoiding or less weighting, the erroneous frames.



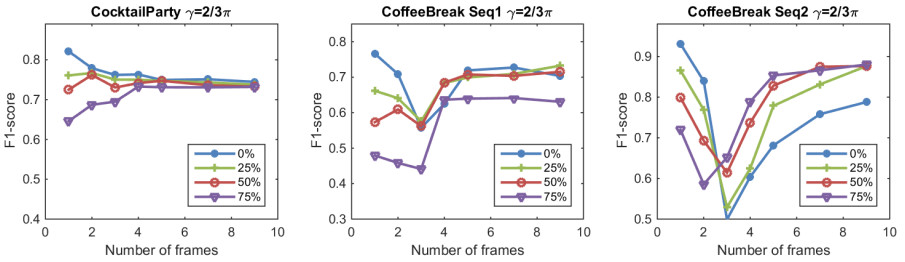
(a)



(b)



(c)



(d)

Figure 9: The results using the different heuristics: a) Normalized sum of weights (Sec. 4.1.1). b) Weighted solutions (Sec. 4.1.2). c) Maximal Entropy (Sec. 4.1.3). d) Clustering Ensemble (Sec. 4.1.4).

Fig. 9b . The weighted solution using the characteristic vector is an intriguing alternative to the normalized sum. The first dominant set extracted in fact captures the peculiarity of the entire graph assigning a high score to the most similar node. This means that the set of weights which share the larger similarity with respect to the other possible solutions, will have a higher score. Statistically, the noisy part should be less than the entire set, hence the set of weights with the highest score will have a good chance to be the right one.

Fig. 9c . The use of the maximal entropy is motivated by the fact that when the entropy is maximized a uniform distribution is obtained. In our case, we have an equal weight distribution only in the case when all the frames are exactly the same, but in most of the cases this does not happen. For this reason, searching for the maximal entropy means to find the set of weights giving a chance to all frames to participate in the final grouping process without suppressing the ability of the optimizer in assigning more weight to the most diverse matrices.

Fig. 9d . The performance of the ensemble of clusters starts from very low results that rapidly grow as the number of frames increases. This is motivated by the fact that when few frames are taken into account (2..4) the number of weights that the algorithm in Sec.4.1 finds is very few. This leads to very different clustering results in which it is difficult to find a consensus. When the number of frames increases, so do the weights, and thus finding a consensus becomes an easier task. For example, with 9 frames the average number of possible weights are more than 150, this means that we will have more than 150 clustering results in which finding a consensus is quite an easy task.

As one can note, in general, there is a performance drop in correspondence of $K = 5$: this is quite obvious because after 5 frames the changes in the scene starts being consistent. This occurs in particular in Seq 1 of the CocktailParty dataset, because in that sequence the groups change frequently. To achieve a good smoothing is thus suggested to use no more than 5 consecutive frames.

Compared with the single-frame approach, in a noiseless tracking situation (blue curve), this version gives comparable results. As shown in the Fig. 9, the temporal

integration varies almost uniformly except a slight increase in the Seq1 of the CoffeeBreak dataset. 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 in general completely dominated by the multi-frame version, irrespective of the number of frames considered in the buffer.

For the sake of curiosity we carried out a last experiment with the aim of establishing if the existing relationship between speed and visual field of view, the so-called *tunnelling* effect affecting the visual field of view on high speed moving, could be confirmed in scenarios in which the velocity of persons are not as high as in cars. To this end, we set the following experiment: for the three datasets in the multi-frame experiment we changed the length and the aperture of the frustum (l and θ parameter) and we correlate them with the speed typical of the dataset used. We end up that there were no relationship at all.

5.5. Discussion

After this empirical evidence we can provide an overall final analysis. The proposed approach is to be preferred over the others under a wide variety of different scenarios. In general, the performance is incredibly stable under both noisy (real) and ideal (synthetic) dataset. For example, we have the highest performance in the CoffeeBreak even if it is a very noisy dataset in terms of head orientation since only 4 orientations are possible. From the single-frame experiments, it is clear that the JS 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 noisy multi- and the single-frame results reveals the meaningfulness of considering consecutive instants of the same scene to to smooth out the noise effect. From the computational point of view, the multi-frame approach is to be preferred in case of noisy measures, using a window of no more than 5 frames. If more frames are considered, the found solutions could be inconsistent. If near real-time detection is required, the single-frame approach is to be preferred over the multiple frame, because it is able to perform group detection at 15/20 fps. Summa-

rizing, the more significant processing modules that absolutely contributed the most in this work and that represent the main novelty, are the biologically inspired model of the frustum, which capture far better the sociological interaction between individuals 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 frames.

As very final note, we analyze the timing requirements for the two approaches, the single- and multi-frame modalities. In the former case, the time requirement is very low, reaching the 15/20 fps (given the detections). Things get worse in the latter case, since the fastest heuristics represented by the *weighted summation* and the *maximal entropy* weights (Sec.4.1.14.1.3) reaches peaks of 2-5 fps (depending by the number of frames). This is quite obvious due to the fact that after having found the set of weights, the final re-weighting is based on a simple sum or sum of logarithms. In second position, we find the *weighted solutions* (Sec.4.1.2), in which a Dominant Set extraction is performed over the set of weights to find the most representative ones. In the last position, we have the *clustering ensemble* (Sec.4.1.4) in which we perform a separate clustering on each set of weights and the final solution is found again by clustering over the evidence accumulator matrix.

The code has been written in (non optimized) Matlab on a Core i7-3720QM2.60GHz with 8GB of RAM.

6. Conclusions

In this paper, we have proposed a method for detecting conversational groups (F-Formations) that can be included in a typical surveillance pipeline or on top of a person detector. The approach improves upon existing methods by building a stochastic model of social attention which captures pairwise scores between people, indicating their joint tendency in aggregating in a group. Pairwise scores fill an affinity matrix which encodes an edge weighted graph representing the entire scene under analsysis. On this structure, a game-theoretic clustering strategy efficiently finds the groups. In addition, this game-theoretic perspective has allowed us to integrate in a principled

way information coming from multiple consecutive frames in videos, in an attempt to deal with noisy situations resulting from the scene complexity (e.g., a crowded high density scenario) and the inaccuracy of the detection and orientation estimation algorithms. Our extensive experimental session on single-frame situation has shown a dramatic improvement over other methods in the literature on five different datasets, and competitive performances on other two benchmarks. Adding the integration with multiple-frames, where applicable, has allowed to augment the overall group detection accuracy, especially in the case of strong noise altering person positions and the related head orientations. In the future, we plan to address the problem of modeling F-formations by considering the instability points, that is, when a group is forming or disgregating, with the challenge of guessing as soon as possible when a person will join or leave a group.

7. Acknowledgements

Authors want to acknowledge Wongun Choi and Yu-Wei Chao to provides us the code, the data of their paper and the valuable support. Hayley Hung has been partially supported by the European Commission under contract number FP7-ICT-600877 (SPENCER) and is affiliated with the Delft Data Science consortium.

References

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

[2] R. Li, P. Porfilio, T. Zickler, Finding group interactions in social clutter, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[3] T. Gan, Y. Wong, D. Zhang, M. S. Kankanhalli, Temporal encoded F-formation system for social interaction detection, in: Proceedings of the 21st ACM international conference on Multimedia, MM '13, ACM, New York, NY, USA, 2013, pp. 937–946. doi:10.1145/2502081.2502096.
URL http://doi.acm.org/10.1145/2502081.2502096

[4] M. Marin-Jimenez, A. Zisserman, V. Ferrari, Here's looking at you, kid. detecting people looking at each other in videos, in: British Machine Vision Conference, 2011.

[5] G. Zen, B. Lepri, E. Ricci, O. Lanz, Space speaks: towards socially and personality aware visual surveillance, in: 1st ACM international workshop on Multimodal pervasive video analysis, 2010, pp. 37–42. doi:10.1145/1878039.1878048.
URL http://doi.acm.org/10.1145/1878039.1878048

[6] H. Hung, B. Kröse, Detecting F-formations as dominant sets, in: ICMI, 2011.

[7] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, V. Murino, Social interaction discovery by statistical analysis of F-formations, in: Proc. of BMVC, BMVA Press, 2011, pp. 23.1–23.12.

[8] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, G. Mori, Discriminative Latent Models for Recognizing Contextual Group Activities, IEEE Trans. Pattern Anal. Mach. Intell. 34 (8) (2012) 1549–1562. doi:10.1109/TPAMI.2011.228.
URL http://dx.doi.org/10.1109/TPAMI.2011.228

[9] T. Yu, S. Lim, K. A. Patwardhan, N. Krahnstoever, Monitoring, Recognizing and Discovering Social Networks, in: CVPR, 2009.

[10] K. Tran, A. Gala, I. Kakadiaris, S. Shah, Activity Analysis in Crowded Environments Using Social Cues for Group Discovery and Human Interaction Modeling, Pattern Recognition Letters.

[11] H. Garfinkel, Studies in Ethnomethodology, Prentice-Hall, 1967.

[12] D. Schweingruber, C. McPhail, A method for systematically observing and recording collective action, Sociological methods & research 27 (4) (1999) 451–498.

[13] F. Setti, C. Russell, C. Bassetti, M. Cristani, F-formation detection: Individuating free-standing conversational groups in images, PloS one 10 (5).

[14] A. Kendon, Conducting Interaction: Patterns of Behavior in Focused Encounters (Studies in Interactional Sociolinguistics), Cambridge University Press, 1990.

[15] H. Hüttenrauch, K. S. Eklundh, A. Green, E. A. Topp, Investigating spatial relationships in human-robot interaction, in: Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on, IEEE, 2006, pp. 5052–5059.

[16] W. Choi, S. Savarese, A unified framework for multi-target tracking and collective activity recognition, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 215–230.

[17] E. T. Hall, The Hidden Dimension, Anchor, 1990.

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

[19] T. M. Ciolek, A. Kendon, Environment and the Spatial Arrangement of Conversational Encounters, Sociological Inquiry 50 (3-4) (1980) 237–271.

[20] C. Chen, J. Odobez, 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, pp. 1544–1551.

[21] V. Jain, J. L. Crowley, Head pose estimation using multi-scale gaussian derivatives, in: Image Analysis, Springer, 2013, pp. 319–328.

[22] F. Setti, H. Hung, M. Cristani, Group Detection in Still Images by F-formation Modeling: a Comparative Study, in: WIAMIS, 2013.

[23] A. Torsello, S. Rota Bulò, M. Pelillo, Grouping with asymmetric affinities: A game-theoretic perspective, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, 2006, pp. 292–299.

[24] S. Rota Bulò, M. Pelillo, A game-theoretic approach to hypergraph clustering, IEEE Trans. Pattern Anal. Machine Intell. 35 (6) (2013) 1312–1327.

[25] K. Somasundaram, J. S. Baras, Achieving symmetric Pareto Nash equilibria using biased replicator dynamics., in: 48th IEEE Conf. Decision Control, 2009, pp. 7000–7005.

[26] S. Pellegrini, A. Ess, L. V. Gool, Improving data association by joint modeling of pedestrian trajectories and groupings, in: European Conference on Computer Vision (ECCV), 2010, pp. 452–465.

[27] K. Yamaguchi, A. Berg, L. Ortiz, T. Berg, Who are you with and where are you going?, in: IEEE Conference on Computer Vision and Patter Recognition (CVPR), 2011.

[28] W. Ge, R. T. Collins, R. B. Ruback, Vision-based analysis of small groups in pedestrian crowds, IEEE Trans. on Pattern Analysis and Machine Intelligence 34 (5) (2012) 1003–1016.

[29] Z. Qin, C. R. Shelton, Improving multi-target tracking via social grouping, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[30] M. Chang, N. Krahnstoever, W. Ge, Probabilistic group-level motion analysis and scenario recognition, in: IEEE ICCV, 2011.

[31] L. Leal-Taixé, G. Pons-Moll, B. Rosenhahn, 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.

[32] S. J. Mckenna, S. Jabri, Z. Duric, H. Wechsler, A. Rosenfeld, Tracking groups of people, Computer Vision and Image Understanding.

[33] F. Cupillard, F. Brémond, M. Thonnat, I. S. Antipolis, O. Group, Tracking groups of people for video surveillance, in: University of Kingston (London), 2001.

[34] J. S. Marques, P. M. Jorge, A. J. Abrantes, J. M. Lemos, Tracking groups of pedestrians in video sequences, in: IEEE Conference on Computer Vision and Patter Recognition Workshops (CVPR workshops), Vol. 9, 2003, pp. 101–101. doi:10.1109/CVPRW.2003.10103.

[35] W. Choi, Y. W. Chao, C. Pantofaru, S. Savarese, Discovering groups of people in images, in: ECCV, 2014.

[36] R. Mora-Colque, G. Cámara-Chávez, W. Schwartz, Detection of Groups of People in Surveillance Videos Based on Spatio-Temporal Clues, in: E. Bayro-Corrochano, E. Hancock (Eds.), Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Vol. 8827 of Lecture Notes in Computer Science, Springer International Publishing, 2014, pp. 948–955. doi:10.1007/978-3-319-12568-8_115.
URL http://dx.doi.org/10.1007/978-3-319-12568-8_115

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

[38] D. Helbing, P. Molnar, Social force model for pedestrian dynamics, Physical review E 51 (5) (1995) 4282.

[39] K. Smith, S. O. Ba, J.-M. Odobez, D. Gatica-Perez, Tracking the Visual Focus of Attention for a Varying Number of Wandering People, IEEE Transactions on

Pattern Analysis and Machine Intelligence 30 (7) (2008) 1212–1229. doi:10.1109/TPAMI.2007.70773.

[40] R. B. Adams, The science of social vision, Vol. 7, Oxford University Press, 2011.

[41] A. Kendon, Some functions of gaze-direction in social interaction., Acta Psychol (Amst) 26 (1) (1967) 22–63.
URL http://view.ncbi.nlm.nih.gov/pubmed/6043092

[42] N. Jovanovic, R. op den Akker, Towards automatic addressee identification in multi-party dialogues, in: Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue, Association for Computational Linguistics, Pennsylvania, USA, 2004, pp. 89–92, imported from HMI.
URL http://doc.utwente.nl/66324

[43] S. Duncan, Some signals and rules for taking speaking turns in conversations, Journal of Personality and Social Psychology 23 (1972) 283–292.

[44] S. O. Ba, J. Odobez, Multiperson visual focus of attention from head pose and meeting contextual cues, Pattern Analysis and Machine Intelligence, IEEE Transactions on 33 (1) (2011) 101–116.

[45] M. Pavan, M. Pelillo, Dominant sets and pairwise clustering, IEEE Trans. Pattern Anal. Machine Intell. 29 (1) (2007) 167–172.

[46] F. Setti, O. Lanz, R. Ferrario, V. Murino, M. Cristani, Multi-Scale F-Formation Discovery for Group Detection, in: International Conference on Image Processing (ICIP), 2013.

[47] S. Vascon, Z. E. Mequanint, M. Cristani, H. Hung, M. Pelillo, V. Murino, A game-theoretic probabilistic approach for detecting conversational groups, in: Proceedings, Asian Conference on Computer Vision (ACCV), LNCS, Springer, Heidelberg, Germany, 2014.

[48] J. Lin, Divergence measures based on the shannon entropy, IEEE Transactions on Information theory 37 (1991) 145–151.

[49]

[50] J. W. Weibull, Evolutionary Game Theory, MIT Press, Cambridge, MA, 2005.

[51] D. Blackwell, An analog of the minimax theorem for vector payoffs, Pacific J. Math. 6 (11) (1956) 1–8.

[52] L. S. Shapley, Equilibrium points in games with vector payoffs, Naval Res. Logistics Quarterly 6 (1959) 57–61.

[53] B. M. Contini, A decision model under uncertainty with multiple objectives, in: A. Mensch (Ed.), Theory of Games: Techniques and Applications, New York, 1966.

[54] M. Zeleny, Games with multiple payoffs, Int. J. Game Theory 4 (4) (1975) 179–191.

[55] T. Kawamura, T. Kanazawa, T. Ushio, Evolutionarily and neutrally stable strategies in multicriteria games, IEICE Trans. Fundam. Electr. Commun. Comp. Sci. E96-A (4) (2013) 814–820.

[56] M. Ehrgott, Multicriteria Optimization, Springer, Berlin, 2005, 2nd edition.

[57] A. Fred, A. Jain, Data clustering using evidence accumulation, in: Pattern Recognition, 2002. Proceedings. 16th International Conference on, Vol. 4, 2002, pp. 276–280 vol.4. doi:10.1109/ICPR.2002.1047450.

[58] L. Bazzani, V. Murino, M. Cristani, Decentralized particle filter for joint individual-group tracking, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[59] E. T. Hall, The Hidden Dimension, Doubleday, Garden City, NY, USA, 1966.

[60] M. Cristani, G. Paggetti, A. Vinciarelli, L. Bazzani, G. Menegaz, V. Murino, Towards computational proxemics: Inferring social relations from interpersonal distances, in: SocialCom/PASSAT 2011, 2011, pp. 290–297.

[61] L. Bazzani, M. Cristani, D. Tosato, M. Farenzena, G. Paggetti, G. Menegaz, V. Murino, Social interactions by visual focus of attention in a three-dimensional environment, Expert Systems 30 (2) (2013) 115–127. doi:10.1111/j.1468-0394.2012.00622.x.
URL http://dx.doi.org/10.1111/j.1468-0394.2012.00622.x

Sebastiano Vascon is currently a PhD student at the Pattern Analysis and Computer Vision department at the Istituto Italiano di Tecnologia of Genova. He studied Computer Science at the University Ca' Foscari of Venice where he received the BSc degree in 2009 and the MSc degree cum laude in 2012 under the supervision of prof. Pelillo and prof. Torsello. During the MSc he spent 6 months at the University College of London under the Erasmus project. His main interests are on pattern recognition, computer vision, graph-theory and game-theory with applications in clustering, medical imaging, scene understanding and behavior analysis.

Eyasu Zemene Mequanint received the BSc degree in Electrical Engineering in 2007 from Jimma University, has worked in Ethio Telecom for 4 years till he has joined CaFoscari University (October 2011) where he got his MSc in Computer Science in June 2013, he has then won a 1 year research fellow to work on Adversarial Learning at Pattern Recognition and Application lab of university of Cagliari. Since September 2014 he is a PhD student of Ca'Foscari University under the supervision of prof. Pelillo. His research interests include: Computer Vision, Multiobjetive Optimization , Game Theoretic models, Machine Learning and Bioinformatics.

Cristani Marco (Ph.D.) is associate professor since 2014 at the Universit degli Studi di Verona, Department of Computer Science, where he teaches and does research within the VIsion, Processing and Sound lab (VIPS). He is also Associate Member of the National Research Council (CNR) and Research Affiliate with the Istituto Italiano di Tecnologia, Genova, Italy, where he was Team Leader since 2009-12. His interests are focused on generative modeling and in particular on generative embeddings, with applications on social signal processing and multimedia.

Hayley Hung is an Assistant Professor and Delft Technology Fellow in the Pattern Recognition and Bioinformatics group at TU Delft, The Netherlands, since 2013. Between 2010-2013, she held a Marie Curie Intra-European Fellowship at the Intelligent Systems Lab at the University of Amsterdam. Between 2007-2010, she was a post-doctoral researcher at Idiap Research Institute in Switzerland. She obtained her PhD in Computer Vision from Queen Mary University of London, UK in 2007 and her first degree from Imperial College, UK in Electrical and Electronic Engineering. Her research interests are in social computing, social signal processing, computer vision, and machine learning.

Marcello Pelillo is Full Professor of Computer Science at Ca Foscari University in Venice, Italy, where he directs the European Centre for Living Technology (ECLT) and leads the Computer Vision and Pattern Recognition group. He held visiting research positions at Yale University, McGill University, the University of Vienna, York University (UK), the University College London, and the National ICT Australia (NICTA). He has published more than 200 technical papers in refereed journals, handbooks, and conference proceedings in the areas of pattern recognition, computer vision and machine learning. He is General Chair for ICCV 2017 and has served as Program Chair for several conferences and workshops (EMMCVPR, SIMBAD, S+SSPR, etc.). He serves (has served) on the Editorial Boards of the journals IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Pattern Recognition, IET Computer Vision, Frontiers in Computer Image Analysis, Brain Informatics, and serves on the Advisory Board of the International Journal of Machine Learning and Cybernetics. Prof. Pelillo is a Fellow of the IEEE and a Fellow of the IAPR.

Vittorio Murino is full professor and director of the PAVIS (Pattern Analysis and Computer Vision) department at the Istituto Italiano di Tecnologia, Genova, Italy. He received the Ph.D. in Electronic Engineering and Computer Science in 1993 at the University of Genova, Italy. Then, he was first at the University of Udine and, since 1998, at the University of Verona, where he served as chairman of the Department of Computer Science from 2001 to 2007. His research interests are in computer vision and machine learning, in particular, probabilistic techniques for image and video processing, with applications on video surveillance, biomedical image analysis and bioinformatics. He is IEEE Senior Member and IAPR Fellow.