

# Groups and Crowds: Behaviour Analysis of People Aggregations

Sadegh Mohammadi<sup>1</sup>(✉), Francesco Setti<sup>2</sup>, Alessandro Perina<sup>3</sup>,  
Marco Cristani<sup>2</sup>, and Vittorio Murino<sup>1,2</sup>

<sup>1</sup> Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia,  
Genova, Italy

{sadegh.mohammadi,vittorio.murino}@iit.it

<sup>2</sup> Department of Computer Science, University of Verona, Verona, Italy

{francesco.setti,marco.cristani,vittorio.murino}@univar.it

<sup>3</sup> Microsoft Corp, WDG Core Data Science, Redmond, USA  
alperina@microsoft.com

**Abstract.** Automatic analysis of human behavior in social environment is a key topic for the computer vision community, with applications in security and video surveillance. While human behavior at an individual (single person) level has been widely studied in the past years, analysis of groups and crowd behavior, is still at a preliminary stage, with room for new approaches to emerge. Recently, there has been significant research effort dedicated to the development of automated computer vision techniques, intended to enhance safety of our societies by monitoring human behaviors and their actions in *groups* and *crowd* level. In particular, groups are usually formed by number of people who gathered for private meeting, birthday party, or wedding, while we consider crowd as huge number of people are gathered together to participate for a national or religious event, or protest due to some dissatisfaction. In this chapter, we will provide a broad overview on proposed approaches on human behavior analysis in group and crowd level, as well as, a detailed of some most recent state-of-the-art methods along with extensive experiments and comparison.

## 1 Introduction

Analyzing the visual content of scenes in videos is increasingly becoming an active research area in computer vision, due to its growing demand in security and surveillance applications. The content of a video captured by surveillance cameras can be potentially monitored by expert personnel for retaining public safety and reducing social crimes in crowded places such as airports, stadiums and malls. However, this is drastically limited by the scarcity of trained personnel and the natural limitation of human attention capabilities to monitor a huge amount of videos simultaneously filmed by multiple surveillance cameras [21]. This hurdle has motivated vision communities to develop methods for automated analysis of crowd scenes recorded by surveillance cameras.

Hitherto, numerous computer vision techniques have been successfully developed to detect and understand human activities in video data [45]. These techniques could be divided in two fundamental types: those which understand the social dynamics occurring in the scene by lying on the fine analysis of each single individual in terms of fine-grained cues (gestures, head pose orientation, feet orientation etc.), and those that operate on more crowded scenarios, where the number of individuals is so high that a per-pedestrian robust analysis cannot be envisaged. In the first case, one of the most intriguing analysis is the number of people who formed the groups within a scene and their activities. *Detection of groups of interacting people* is a very interesting and useful task in many modern technologies, with application fields spanning from video-surveillance to social robotics.

In the case that per-person fine analysis cannot be carried out, another branch of approaches should be taken into account. For example, in the case of highly crowded scenarios (more than 30 people), person detection, tracking, gesture recognition and other techniques for fine-grained analysis are often degraded by the presence of severe occlusions, cluttered background, low quality of surveillance data and, most importantly, by the complex interplays among people involved in crowd [28].

That has opened up a new broad research line which is generally referred to as *crowd scene analysis* in the computer vision literature [33,64].

In this chapter, we analyze the very last state of the art related to the group analysis, together with the latest results in terms of crowd scene analysis.

A group can be broadly understood as a social unit comprising several members who stand in status and relationships with one another [5]. However, there are many kinds of groups, that differ in dimension (small groups or crowds), durability (ephemeral, ad hoc or stable groups), in/formality of organization, degree of sense of belonging, level of physical dispersion, etc. [6] (see the literature review in the next section). In this article, we build from the concepts of sociological analysis and we focus on free-standing conversational groups (FCGs), or small ensembles of co-present persons engaged in ad hoc focused encounters [6–8]. FCGs represent crucial social situations, and one of the most fundamental bases of dynamic sociality: these facts make them a crucial target for the modern automated monitoring and profiling strategies which have started to appear in the literature in the last three years [3,9–14]. FCGs emerge during many and diverse social occasions, such as a party, a social dinner, a coffee break, a visit in a museum, a day at the seaside, a walk in the city plaza or at the mall; more generally, when people spontaneously decide to be in each others immediate presence to interact with one another. For these reasons, FCGs are fundamental social entities, whose automated analysis may bring to a novel level of activity and behavior analysis.

In a FCG, people communicate to the other participants, among and above all the rest, what they think they are doing together, what they regard as the activity at hand. And they do so not only, and perhaps not so much, by talking, but also, and as much, by exploiting non-verbal modalities of expression, also

called social signals [23], among which positional and orientational forms play a crucial role (cf. also [7], p. 11). In fact, the spatial position and orientation of people define one of the most important proxemic notions which describe an FCG, that is, Adam Kendons Facing Formation, mostly known as F-formation.

Detecting free-standing conversational groups is useful in many contexts. In video-surveillance, automatically understanding the network of social relationships observed in an ecological scenario may result beneficial for advanced suspect profiling, improving and automatizing SPOT (Screening Passengers by Observation Technique) protocols [26], which nowadays are performed uniquely by human operators. In this chapter we analyze one of the latest technique for group detection, acting on single images acquired by a monocular camera, which operates on positional and orientational information of the individuals in the scene. Unlike previous approaches, the methodology is a direct formulation of the sociological principles (proximity, orientation and ease of access) concerning F-formations.

However, the aforementioned approaches is useful for moderate crowd scenarios, where the people are segmentable, and we can track them within a frame of video. Whilst, this is not a case for crowd scenes. Therefore, crowd scene analysis has recently attracted intense attention from the vision community. In particular, proposed method in this area can be categorized into three topics, including (1) crowd density estimation and people counting, (2) tracking in crowd, and (3) modeling crowd behaviors [21]. Estimating the number of people in a crowd is the foremost stage for several real-world applications such as safety control, monitoring public transportation, crowd rendering for animation and crowd simulation for urban planning. Despite many significant works in this area [13, 17], automated crowd density estimation still remains an open problem in computer vision due to extreme occlusions and visual ambiguities of human appearance in crowd images [50]. Tracking individuals (or objects) in crowd scenes is another challenging task [48, 56]: other than severe occlusions, cluttered background and pattern deformations, which are common difficulties in visual object tracking, the efficiency of crowd trackers is largely dependent on crowd density and dynamics, people social interactions as well as the crowd’s psychological characteristics [2]<sup>1</sup>.

The primary goal of modeling crowd behaviors is to identify abnormal events such as riot, panic and violence in crowd scenes [29]. Despite recent success in this research field, detecting crowd abnormalities still remain an open and very challenging problem. The biggest issue of crowd anomaly detection lies in the definition of abnormality as it is strongly context dependent [25, 37]. For example, riding a bike in a street is a normal action, whereas it is considered abnormal in another scene with a different context such as a park or sidewalk. Similarly, people gathering for a social event is a normal event, while same gathering at the same place to “protest against a law” is an abnormal event. Another challenge stems from the lack of adequate training samples to learn a well-generalized crowd model. This drastically degrades the generalization power of current crowd

---

<sup>1</sup> Readers are referred to [33, 50] for a full treatment on the tasks of crowd tracking and density estimation.

models, since they are not capable of capturing the large intra-class variations of crowd behaviors [21].

Concerning the crowd scene analysis, in this chapter we will overview some leading techniques in the computer vision literature designed for detecting abnormal behaviors in crowd, with a focus on existing *motion* and *model* based approaches. Then, we will give a general overview on the most recent approaches. Finally, we will extensively evaluate their performance on various challenging imaging and crowding conditions.

The rest of the paper will be organized as follows: the first part is related to the analysis of the group, in the case single individuals can be captured and modeled. In the second part we will consider the approaches of crowd scene analysis.

## 2 Groups: Related Work

In computer vision, the analysis of groups has occurred historically in two broad contexts: video-surveillance and meeting analysis.

Within the scope of video-surveillance, the definition of a group is generally simplified to two or more people of similar velocity, spatially and temporally close to one another [15]. This simplified definition arises from the difficulty of inferring persistent social structure from short video clips. In this case, most of the vision-based approaches perform group tracking, i.e. capturing individuals in movement and maintaining their identity across video frames, understanding how they are partitioned in groups [4, 15–19].

In meeting analysis, typified by classroom behavior [1], people usually sit around a table and remain near a fixed location for most of the time, predominantly interacting through speech and gesture. In such a scenario, activities can be finely monitored using a variety of audiovisual features, captured by pervasive sensors like portable devices, microphone arrays, etc. [20–22].

From a sociological point of view, meetings are examples of social organization that employs focused interaction, which occurs when persons openly cooperate to sustain a single focus of attention [6, 7]. This broad definition covers other collaborative situated systems of activity that entail a more or less static spatial and proxemic organization such as playing a board or sport game, having dinner, doing a puzzle together, pitching a tent, or free conversation [6], whether sitting on the couch at a friends place, standing in the foyer and discussing the movie, or leaning on the balcony and smoking a cigarette during work-break.

Free-standing conversational groups (FCGs) [8] are another example of focused encounters. FCGs emerge during many and diverse social occasions, such as a party, a social dinner, a coffee break, a visit in a museum, a day at the seaside, a walk in the city plaza or at the mall; more generally, when people spontaneously decide to be in each others immediate presence to interact with one another. For these reasons, FCGs are fundamental social entities, whose automated analysis may bring to a novel level of activity and behavior analysis.

A robust FCG detector may also impact the social robotics field, where the approaches so far implemented work on few number of people, usually focusing on a single F-formation [27–29].

Efficient identification of FCGs could be of use in multimedia applications like mobile visual search [30,31], and especially in semantic tagging [32,33], where groups of people are currently inferred by the proximity of their faces in the image plane. Adopting systems for 3D pose estimation from 2D images [34] plus an FCG detector could in principle lead to more robust estimations. In this scenario, the extraction of social relationships could help in inferring personality traits [35,36] and triggering friendship invitation mechanisms [37].

In computer-supported cooperative work (CSCW), being capable of automatically detecting FCGs could be a step ahead in understanding how computer systems can support socialization and collaborative activities: e.g., [38–41]; in this case, FCGs are usually found by hand, or employing wearable sensors.

Manual detection of FCGs occurs also in human computer interaction, for the design of devices reacting to a situational change [42,43]: here the benefit of the automation of the detection process may lead to a genuine systematic study of how proxemic factors shape the usability of the device.

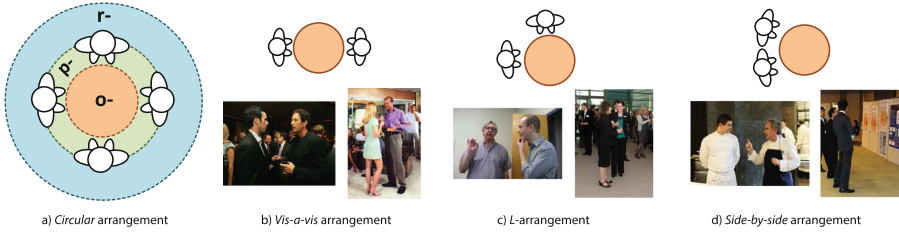
The last three years have seen works that automatically detect F-formations: Bazzani et al. [9] first proposed the use of positional and orientational information to capture Steady Conversational Groups (SCG); Cristani et al. [3] designed a sampling technique to seek F-formations centres by performing a greedy maximization in a Hough voting space; Hung and Kröse [10] detected F-formations by finding distinct maximal cliques in weighted graphs via graph-theoretic clustering; both the techniques were compared by Setti et al. [12]. A multi-scale extension of the Hough-based approach [3] was proposed by Setti et al. [13]. This improved on previous works, by explicitly modeling F-formations of different cardinalities. Tran et al. [14] followed the graph based approach of [10], extending it to deal with video-sequences and recognizing five kinds of activities. Vascon *et al.* [60] employed a games-theoretic approach to deal with dominant sets in order to detect stati F-formations. Lastly, in [53] Setti *et al.* proposed a graph-cut technique that outperformed all the previous methods. In the following sections we will detail this method and present experimental results comparing all the above mentioned algorithms.

### 3 Graph-Cuts for F-Formation

GCFF method is strongly based on the formal definition of F-formation given by Kendon [27] (*page 209*):

*An F-formation arises whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access.*

According to this definition, an F-formation is the proper organisation of three social spaces: *o-space*, *p-space* and *r-space* (see Fig. 1a). The *o-space* is a



**Fig. 1.** Structure of an F-formation and examples of F-formation arrangements. (a) Schematization of the three spaces of an F-formation: starting from the centre, *o-space*, *p-space* and *r-space*. (b–d) Three examples of F-formation arrangements: for each one of them, one picture highlights the head and shoulder pose, the other shows the lower body posture.

convex empty space surrounded by the people involved in a social interaction, where every participant is oriented inward into it, and no external people are allowed to lie. More in the detail, the *o-space* is determined by the overlap of those regions dubbed *transactional segments*, where as transactional segment we refer to the area in front of the body that can be reached easily, and where hearing and sight are most effective [15]. In practice, in a F-formation, the transactional segment of a person coincides with the *o-space*, and this fact has been exploited in our algorithm. The *p-space* is the belt of space enveloping the *o-space*, where only the bodies of the F-formation participants (as well as some of their belongings) are placed. People in the *p-space* participate to an F-formation using the *o-space* to transmit their messages. The *r-space* is the space enveloping *o-* and *p-spaces*, and is also monitored by the F-formation participants. People joining or leaving a given F-formation mark their arrival as well as their departure by engaging in special behaviours displayed in a special order in special portions of *r-space*, depending on several factors (context, culture, personality among the others); therefore, here we prefer to avoid the analysis of such complex dynamics, leaving their computational analysis as future work.

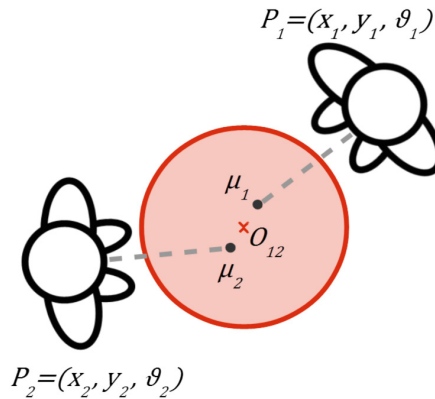
F-formations can be organised in different *arrangements*, that is, spatial and orientational layouts (see Fig. 1a–d) [16, 18, 27]. In F-formations of two individuals, usually we have a *vis-a-vis* arrangement, in which the two participants stand and face one another directly; another situation is the *L-arrangement*, when two people lie in a right angle to each other. As studied by Kendon [27], *vis-a-vis* configurations are preferred for competitive interactions, whereas L-shaped configurations are associated with cooperative interactions. In a *side-by-side* arrangement, people stand close together, both facing the same way; this situation occurs frequently when people stand at the edges of a setting against walls. *Circular* arrangements, finally, hold when F-formations are composed by more than two people; other than being circular, they can assume an approximately linear, semicircular, or rectangular shape.

Graph-Cuts for F-Formation finds the *o-space* of an F-formation, assigning to it those individuals whose transactional segments do overlap, without

focusing on a particular arrangement. Given the position of an individual, to identify the transactional segment we exploit orientational information, which may come from the head orientation, the shoulder orientation or the feet layout, in increasing order of reliability [27]. The idea is that the feet layout of a subject indicates the mean direction along which his messages should be delivered, while he is still free to rotate his head and to some extent his shoulders through a considerable arc, before he must begin to turn his lower body as well. The problem is that feet are almost impossible to detect in an automatic fashion, due to the frequent (auto) occlusions; shoulder orientation is also complicated, since most of the approaches of body pose estimation work on 2D data and do not manage auto-occlusions. However, since any sustained head orientation in a given direction is usually associated with a reorientation of the lower body (so that the direction of the transactional segment again coincides with the direction in which the face is oriented [27]), head orientation should be considered proper for detecting transactional segments and, as a consequence, the o-space of an F-formation. In this work, we assume to have as input both positional information and head orientation; this assumption is reasonable due to the massive presence of robust tracking technologies [6] and head orientation algorithms [3, 14, 55].

In addition to this, we consider soft exclusion constraints: in an o-space, F-formation participants should have *equal, direct and exclusive access*. In other words, if person  $i$  stands between another person  $j$ , and an o-space centre  $O_g$  of the F-formation  $g$ , this should prevent  $j$  from focusing on the o-space, and, as a consequence, from being part of the related F-formation.

In what follows, we formally define the objective function accounting for positional, orientational and exclusion constraints aspects, and show how it can be optimised. Figure 2 gives a graphical idea of the problem formulation.



**Fig. 2.** Schematic representation of the problem formulation. Two individuals facing each other, the gray dot representing the transitional segment centre, the red cross being the o-space centre and the red area the o-space of the F-formation. (Color figure online)

### 3.1 Objective Function

We use  $P_i = [x_i, y_i, \theta_i]$  to represent the position  $x_i, y_i$  and head orientation  $\theta_i$  of the individual  $i \in \{1, \dots, n\}$  in the scene. Let  $TS_i$  be the a priori distribution which models the transactional segment of individual  $i$ . As we explained in the previous section, this segment is coherent with the position and orientation of the head, so we can assume  $TS_i \sim \mathcal{N}(\mu_i, \Sigma_i)$ , where  $\mu_i = [x_{\mu_i}, y_{\mu_i}] = [x_i + D \cos \theta_i, y_i + D \sin \theta_i]$ ,  $\Sigma_i = \sigma \cdot \mathbf{I}$  with  $\mathbf{I}$  the 2D identity matrix, and  $D$  is the distance between the individual  $i$  and the centre of its transactional segment (hereafter called *stride*). The stride parameter  $D$  can be learned by cross-validation, or fixed a priori accounting for social facts. In practice, we assume the transactional segment of a person having a circular shape, which can be thought as superimposed to the o-space of the F-formation she may be part of.

$O_g = [u_g, v_g]$  indicates the position of a candidate o-space centre for F-formation  $g \in \{1, M\}$ , while we use  $G_i$  to refer to the F-formation containing individual  $i$ , considering the F-formation assignment  $G_i = g$  for some  $g$ . The assignment assumes that each individual  $i$  may belong to a single F-formation  $g$  only at any given time, and this is reasonable when we are focusing one a single time, that is, an image. It follows naturally the definition of  $O_{G_i} = [u_{G_i}, v_{G_i}]$ , which represents the position of a candidate o-space centre for an unknown F-formation  $G_i = g$  containing  $i$ . For the sake of mathematical simplicity, we assume that each lone individual not belonging to a gathering can be considered as a *spurious* F-formation.

At this point, we define the likelihood probability of an individual  $i$ 's transitional segment centre  $C_i = [u_i, v_i]$  given the a priori variable  $TS_i$ .

$$\Pr(C_i | TS_i) \propto \exp\left(-\frac{\|C_i - \mu_i\|_2^2}{\sigma^2}\right) \quad (1)$$

$$= \exp\left(-\frac{(u_i - x_{\mu_i})^2 + (v_i - y_{\mu_i})^2}{\sigma^2}\right) \quad (2)$$

Hence, the probability that an individual  $i$  shares an o-space centre  $O_{G_i}$  is given by

$$\Pr(C_i = O_{G_i} | TS_i) \propto \exp\left(-\frac{(u_{G_i} - x_{\mu_i})^2 + (v_{G_i} - y_{\mu_i})^2}{\sigma^2}\right) \quad (3)$$

and the posterior probability of any overall assignment is given by

$$\Pr(C = O_G | TS) \propto \prod_{i \in [1, n]} \exp\left(-\frac{(u_{G_i} - x_{\mu_i})^2 + (v_{G_i} - y_{\mu_i})^2}{\sigma^2}\right) \quad (4)$$

with  $C$  the random variable which models a possible joint location of all the o-space centres,  $O_G$  is one instance of this joint location, and  $TS$  is the position of all the transitional segments of the individuals in the scene.



Clearly, if the number of o-space centres is unconstrained, the maximum a posteriori probability (MAP) occurs when each individual has his own separate o-space centre, generating a *spurious* F-formation formed by a single individual, that is,  $O_{G_i} = TS_i$ . To prevent this from happening, we associate a minimum description length prior (MDL) over the number of o-space centres used. This prior takes the same form as dictated by the Akaike Information Criterion (AIC) [10], linearly penalising the log-likelihood for the number of models used.

$$\Pr(C = O_G | TS) \propto \prod_{i \in [1, n]} \exp\left(-\frac{(u_{G_i} - x_{\mu_i})^2 + (v_{G_i} - y_{\mu_i})^2}{\sigma^2}\right) \cdot \exp(-|O_G|) \quad (5)$$

where  $|O_G|$  is the number of distinct F-formations.

To find the MAP solution, we take the negative log-likelihood and discarding normalising constants, we have the following objective  $J(\cdot)$  in standard form:

$$J(O_G | TS) = \sum_{i \in [1, n]} (u_{G_i} - x_{\mu_i})^2 + (v_{G_i} - y_{\mu_i})^2 + \sigma^{-2}|O_G| \quad (6)$$

As such, this can be seen as optimizing a least-squares error combined with an MDL prior. In principle this could be optimised using a standard technique such as k-means clustering combined with a brute force search over all possible choices of  $k$  to optimise the MDL cost. In practice, k-means frequently gets stuck in local optima and, in fact, using the technique described the least squares component of the error frequently increases, instead of decreasing, as  $k$  increases. Instead we make use of the graph-cut based optimisation described in [30], and widely used in computer vision [9, 11, 35, 63].

In short, we start from an abundance of possible o-space centres, and then we use a hill-climbing optimisation that alternates between assigning individuals to o-space centres using the efficient graph-cut based optimisation [30] that directly minimises the cost (6), and then minimising the least squares component by updating o-space centres to the mean of  $O_g$ , for all the individuals  $\{i\}$  currently assigned to the F-formation. The whole process is iterated until convergence. This approach is similar to the standard k-means algorithm, sharing both the assignment, and averaging step. However, as the graph-cut algorithm selects the number of clusters, we can avoid local minima by initialising with an excess of model proposals. In practice, we start from the previously mentioned trivial solution in which each individual is associated with its own o-space centre, centred on his position.

### 3.2 Visibility Constraints

Finally, we add the natural constraint that people can only join an F-formation if they can see the o-space centres. By allowing other people to occlude the o-space centre, we are able to capture more subtle nuances such as people being crowded out of F-formations or deliberately ostracised. Broadly speaking, an individual is excluded from an F-formation when another individual stands between him and the group centre. Taking  $\theta_{i,j}^g$  as the angle between two individuals about a

**Algorithm 1.** Finding shared focal centres.

---

```

Initialise with  $O_{G_i} = TS_i \quad \forall i \in [1, \dots, n]$ 
old_cost =  $\infty$ 
while  $J(O_G, TS) < \text{old\_cost}$  do
  old_cost  $\leftarrow J(O_G, TS)$ 
  run graph cuts to minimise cost (6)
  for  $\forall g \in [1, \dots, M]$  do
    if  $g$  is not empty then
      update  $O_G \leftarrow \frac{\sum_{i \in g} TS_i}{|g|}$ 
    end if
  end for
end while

```

---

given o-space centre  $O_g$  for which is assumed  $G_i = G_j = g$  and  $d_i^g, d_j^g$  as the distance of  $i$ , or  $j$ , respectively from the o-space centre  $O_g$ , the following cost captures this property:

$$R_{i,j}(g) = \begin{cases} 0 & \text{if } \theta_{i,j}^g \leq \hat{\theta}, \text{ or } d_i^g < d_j^g \\ \exp(K \cos(\theta_{i,j}^g)) \frac{d_i^g - d_j^g}{d_j^g} & \text{otherwise.} \end{cases} \quad (7)$$

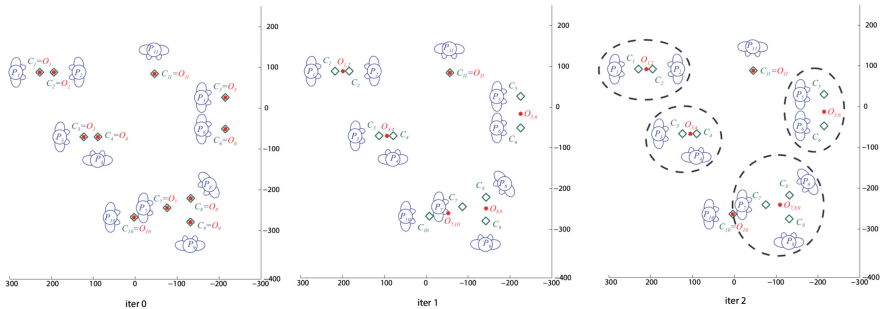
and use the new cost function:

$$J'(O_G|TS) = J(O_G|TS) + \sum_{i,j \in P} R_{i,j}(G_i) \quad (8)$$

$R_{i,j}(g_i)$  acts as a visibility constraint on  $i$  regardless of the group person  $j$  is assigned to, as such it can be treated as a unary cost or data-term and included in the graph-cut based part of the optimisation. Now we turn to other half of the optimisation - updating the o-space centres. Although, given an assignment of people to a o-space centre, a local minima can be found using any off the shelf non-convex optimisation, we take a different approach. There are two points to be aware of: first, the difference between  $J'$  and  $J$  is sharply peaked and close to zero in most locations, and can generally be safely ignored; second and more importantly, we may often want to move out of a local minima. If updating an o-space centre results in a very high repulsion cost to one individual, this can often be dealt with by assigning the individual to a new group, and this will result in a lower overall cost, and more accurate labelling. As such, when optimising the o-space centres, we pass two proposals for each currently active model to graph-cuts – the previous proposal generated, and a new proposal based on the current mean of the F-formation. As the graph-cut based optimisation starts from the previous solution, and only moves to lower cost labellings, the cost always decreases and the procedure is guaranteed to converge to a local optimum.

### 3.3 An Explicative Example

Figure 3 gives a visual insight of our graph-cuts process. Given the position and orientation of each individual  $P_i$ , the algorithm starts by computing the transitional segments  $C_i$ . At the first iteration 0, the candidate o-space centres  $O_i$  are initialized, and are coincident with the transitional segments  $C_i$ ; in this example are present 11 individuals, so 11 candidate o-space centres are generated. After iteration 1, the proposed segmentation process provides 1 singleton ( $P_{11}$ ) and 5 FCGs of two individuals each. We can appreciate different configurations such as *vis-a-vis* ( $O_{1,2}$ ), L-shape ( $O_{3,4}$ ) and side-by-side ( $O_{5,6}$ ). Still, the grouping in the bottom part of the image is wrong ( $P_7$  to  $P_{10}$ ), since it violates the exclusion principle. In iteration 2, the previous candidate o-space centres is considered as initialization, and a new graph is built. In this new configuration, the group  $O_{7,10}$  is recognized as violating the visibility constraint and thus the related edge is penalized; a new run of graph-cuts minimization allows to correctly cluster the FCGs in a singleton ( $P_{10}$ ) and a FCG formed by three individuals ( $O_{7,8,9}$ ), which corresponds to the ground truth (visualized as the dashed circles).



**Fig. 3.** An explicative example. Iteration 0: initialization with the candidate o-space centres  $\{O\}$  coincident with the transitional segment of each individual  $\{C\}$ . Iteration 1: first graph-cuts run; easy groups are correctly clustered while the most complex still present errors (the FCG formed by  $P_7$  and  $P_{20}$  violates the visibility constraint). Iteration 2: the second graph-cuts run correctly detects the  $O_{7,8,9}$  F-formation.

## 4 Groups: Experiments

In this section we present experimental results on five publicly available datasets employed as benchmark, and eight state of the art methods. Here we will show that GCFF definitely outperforms all the competitors, setting in all the cases new state-of-the-art scores. To conclude, we present an extended analysis of how the methods perform in terms of their ability of detecting groups of various cardinality and to test the robustness to noise, further promoting our technique.

Five publicly available datasets are used for the experiments: two from [19] (*Synthetic* and *Coffee Break*), one from [24] (*IDIAP Poster Data*), one from [52]

**Table 1.** Summary of the features of the datasets used for experiments.

Dataset	Data type	Detection	Detection quality	# of frames	# of people	# of people for groups' cardinality					
						1	2	3	4	5	6
Synthetic data	Synthetic	–	Perfect	100	900	300	360	240	–	–	–
IDIAP poster	Real	Manual	Very high	82	1,695	429	910	339	12	5	–
Cocktail party	Real	Automatic	High	320	1,915	174	162	246	176	275	882
Coffee break	Real	Automatic	Low	119	1,299	376	464	459	–	–	–
GDet	Real	Automatic	Very low	403	1,474	367	394	372	88	175	78

(*Cocktail Party*), and one from [5] (*GDet*). A summary of the dataset features is in Table 1, while a detailed presentation of each dataset follows.

**Synthetic Data.** A psychologist generated a set of 10 diverse situations, each one repeated with minor variations for 10 times, resulting in 100 frames representing different social situations, with the aim to span as many configurations as possible for F-formations. An average of 9 individuals and 3 groups are present in the scene, together with some singletons. Proxemic information is noiseless in the sense that there is no clutter in the position and orientation state of each individual.

**IDIAP Poster Data.** Over 3 h of aerial videos (resolution  $654 \times 439$  px) have been recorded during a poster session of a scientific meeting. Over 50 people are walking through the scene, forming several groups over time. A total of 82 images were selected with the idea to maximise the crowdedness and variance of the scenes. Images are unrelated to each other in the sense that there are no consecutive frames, and the time lag between them prevents to exploit temporal smoothness. As for the data annotation, a total of 24 annotators were grouped into 3-person subgroups and they were asked to identify F-formations and their associates from static images. Each person's position and body orientation was manually labelled and recorded as pixel values in the image plane – one pixel represented approximately 1.5 cm.

**Cocktail Party.** This dataset contains about 30 min of video recordings of a cocktail party in a  $30 \text{ m}^2$  lab environment involving 7 subjects. The party was recorded using four synchronised angled-view cameras (15 Hz,  $1024 \times 768$  px, jpeg) installed in the corners of the room. Subjects' positions were logged using a particle filter-based body tracker [31] while head pose estimation is computed as in [32]. Groups in one frame every 5 s were manually annotated by an expert, resulting in a total of 320 labelled frames for evaluation.

**Coffee Break.** The dataset focuses on a coffee-break scenario of a social event, with a maximum of 14 individuals organised in groups of 2 or 3 people each. Images are taken from a single camera with resolution of  $1440 \times 1080$  px. People positions have been estimated by exploiting multi-object tracking on the heads, and head detection has been performed afterwards with the algorithm of [57], considering solely 4 possible orientations (front, back, left and right) in the image

plane. The tracked positions and head orientations were then projected onto the ground plane. Considering the ground truth data, a psychologist annotated the videos indicating the groups present in the scenes, for a total of 119 frames split in two sequences. The annotations were generated by analysing each frame in combination with questionnaires that the subjects filled in.

***GDet.*** The dataset is composed by 5 subsequences of images acquired by 2 angled-view low resolution cameras ( $352 \times 328$  px) with a number of frames spanning from 17 to 132, for a total of 403 annotated frames. The scenario is a vending machines area where people meet and chat while they are having coffee. This is similar to Coffee Break scenario but in this case the scenario is indoor, which makes occlusions many and severe; moreover, people in this scenario knows each other in advance. The videos were acquired with two monocular cameras, located on opposite angles of the room. To ensure the natural behaviour of people involved, they were not aware of the experiment purposes. For ground truth generation, people tracking has been carried out with the particle filter proposed in [31], while head pose estimation is performed afterwards with the method in [57] considering only 4 orientations (front, back, left and right).

We compare the GCF algorithm with seven state of the art methods: one exploiting the concept of *view frustum* (IRPM), three based on dominant-sets (DS, IGD and GTCG), three different version of Hough Voting approaches (linear, entropic and multi-scale HVFF).

***Inter-Relation Pattern Matrix (IRPM).*** Proposed by Bazzani *et al.* [5], uses the head direction to infer the 3D view frustum as approximation of the focus-of-attention of an individual; this is used together with proximity information to estimate interactions: the idea is that close-by people whose view frustum is intersecting are in some way interacting.

***Dominant Sets (DS).*** Presented by Hung and Kröse [24], this algorithm considers an F-formation as a dominant-set cluster [44] of an edge-weighted graph, where each node in the graph is a person, and the edges between them measure the affinity between pairs.

***Interacting Group Discovery (IGD).*** Presented by Tran *et al.* [58], is based on dominant sets extraction from an undirected graph where nodes are individuals and the edges have a weight proportional to how much people are interacting; the attention of an individual is modeled as an ellipse centred at a fixed offset in front of him, while the interaction between two individuals is proportional to the intersection of their attention ellipses.

***Game-Theory for Conversational Groups (GTCG).*** In [59] the authors develop a game-theoretic framework, supported by a statistical modeling of the uncertainty associated with the position and orientation of people. Specifically, they use a representation of the affinity between candidate pairs by expressing the distance between distributions over the most plausible oriented region of attention. Additionally, they can integrate temporal information over multiple frames by using notions from multi-payoff evolutionary game theory.

**Hough Voting for F-Formation (HVFF).** Under this caption we consider a set of methods based on a Hough Voting strategy to build accumulation spaces and find local maxima of this function to identify F-formations. The general idea is that each individual is associated with a Gaussian probability density function which describes the position of the o-space centre he is pointing at. The pdf is approximated by a set of samples, which basically vote for a given o-space centre location. The voting space is then quantized and the votes are aggregated on squared cells, so to form a discrete accumulation space. Local maxima in this space identify o-space centres, and consequently, F-formations. Over the years, three versions of these framework have been presented: in [19] the votes are linearly accumulated by just summing up all the weights of votes belonging to the same cell, in [51] the votes are aggregated by using the weighted Boltzmann entropy function, while in [52] a multi-scale approach is used on top of the entropic version.

As accuracy measures, we adopt the metrics proposed in [19] and then extended in [53]: we consider a group as correctly estimated if at least  $\lceil(T \cdot |G|)\rceil$  of their members are found by the grouping method and correctly detected by the tracker, and if no more than  $1 - \lceil(T \cdot |G|)\rceil$  false subjects (of the detected tracks) are identified, where  $|G|$  is the cardinality of the labelled group  $G$ , and  $T \in ]0, 1]$  is an arbitrary threshold, called *tolerance threshold*. In particular, we focus on two interesting values of  $T$ :  $2/3$  and  $1$ .

With this definition of *tolerant match*, we can determine for each frame the correctly detected groups (true positives – TP), the miss-detected groups (false negatives – FN) and the hallucinated groups (false positives – FP). With this, we compute the standard pattern recognition metrics precision and recall:

$$precision = \frac{TP}{TP + FP}, \quad recall = \frac{TP}{TP + FN} \quad (9)$$

and the  $F_1$  score defined as the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (10)$$

In addition to these metrics, we compute the *Global Tolerant Matching* score (GTM), which is the area under the curve (AUC) in the  $F_1$  vs.  $T$  graph with  $T$  varying from  $1/2$  to  $1$ . Since in our experiments we only have groups up to 6 individuals, without loss of generality we consider  $T$  varying with 3 equal steps in the range stated above.

Moreover, we will discuss results also in terms of group cardinality, by computing the  $F_1$  score for each cardinality separately and then computing mean and standard deviation.

#### 4.1 Best Results Analysis

Given the metrics explained above, the first test analyses the best performances for each method on each dataset; in practice, a tuning phase has been carried

out for each method/dataset combination in order to get the best performances. Note, we did not have code for Dominant Sets [24] and thus we used results provided directly from the authors of the method for a subset of data. For this reason, average results over all the datasets are only averaged over 3 datasets, and cannot be taken into account for a fair comparison. Best parameters are found on half of one sequence by cross-validation, and kept unchanged for the remaining part of the dataset. Please note, finding the right parameters can also be fixed by hand, since the stride  $D$  depends on the social context under analysis (formal meetings will have higher  $D$ , the presence of tables and similar items may also increase the diameter of the FCGs): with a given  $D$ , for example, it is assumed that circular F-formations will have diameter of  $2D$ . The parameter  $\sigma$  indicates how much we are permissive in accepting deviations from such a diameter. Moreover,  $D$  depends also on the different measure units (pixels/cm) which characterize the proxemic information associated to each individual in the scene.

Table 2 shows best results by considering the threshold  $T = 2/3$ , which corresponds to find at least  $2/3$  of the members of a group, no more than  $1/3$  of false subjects; while Table 3 presents results with  $T = 1$ , considering a group as correct if all and only its members are detected. The proposed method outperforms all the competitors, on all the datasets. With  $T = 2/3$ , three observations can be made: the first is that our approach GCOFF improves substantially the precision (of 13% in average) and even more definitely the recall scores (of 17% in average) of the state of the art approaches. The second is that our approach produces the same score for both the precision and the recall; this is very convenient and convincing, since so far all the approaches of FCG detections have shown to be weak in the recall dimension. The third observation is that GCOFF performs well both in the case where no errors in the position or orientation of the people are present (as the Synthetic dataset) and in the cases where strong noise of position and orientation is present (Coffee Break, GDet).

When moving to tolerance threshold equal to 1 (all the people in a group have to be individuated, and no false positive are allowed) the performance is reasonably lower, but the increment is even stronger w.r.t. to the state of the art, in general on all the datasets: in particular, on the Cocktail Party dataset, the results are more than twice the scores of the competitors. Finally, even in this case, GCOFF produces a very similar score for precision and recall.

A performance analysis is also provided by changing the tolerance threshold  $T$ . Figure 4 shows the average  $F_1$  scores for each method computed over all the frames and datasets. From the curves we can appreciate how the proposed method is consistently best performing for each  $T$ -value. In the legend of Fig. 4 the Global Tolerant Matching score is also reported. Again, GCOFF is outperforming the state of the art, independently from the choice of  $T$ .

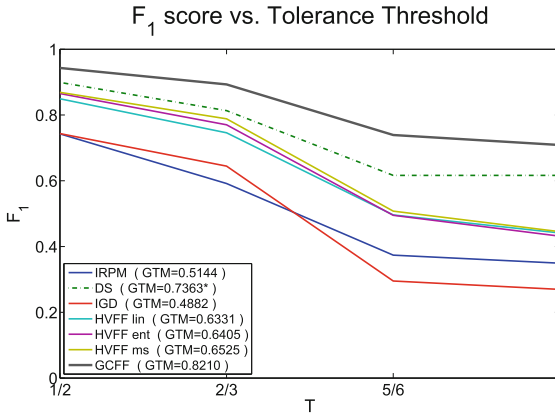
The reason why our approach does better than the competitors has been explained in the state of the art section, here briefly summarized: the Dominant Set-based approaches DS and IGD, even if they are based on an elegant optimization procedure, tend to find circular groups, and are weaker in individuating

**Table 2.** Average precision, recall and  $F_1$  scores for all the methods and all the datasets ( $T = 2/3$ ).

	Synthetic			IDIAP poster			Cocktail party			Coffee break			GDet			Total		
	Prec	Rec	$F_1$	Prec	Rec	$F_1$	Prec	Rec	$F_1$	Prec	Rec	$F_1$	Prec	Rec	$F_1$	Prec	Rec	$F_1$
IRPM [5]	0.85	0.80	0.82	0.82	0.74	0.78	0.56	0.43	0.49	0.68	0.50	0.57	0.77	0.47	0.58	0.70	0.49	0.56
DS [24]	0.85	0.97	0.90	0.91	0.92	0.91	–	–	–	0.69	0.65	0.67	–	–	–	0.81	0.83	0.82
IGD [58]	0.95	0.71	0.81	0.80	0.68	0.73	0.81	0.61	0.70	0.81	0.78	0.79	0.83	0.36	0.50	0.68	0.76	0.70
CTCG [59]	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	0.92	0.96	0.94	<b>0.86</b>	0.82	0.84	0.83	0.89	0.86	0.76	0.76	0.76	0.83	0.83	0.83
HVFF lin [19]	0.75	0.86	0.80	0.90	0.95	0.92	0.59	0.74	0.65	0.73	0.86	0.79	0.66	0.68	0.67	0.75	0.79	0.76
HVFF ent [51]	0.79	0.86	0.82	0.86	0.89	0.87	0.78	0.83	0.80	0.76	0.86	0.81	0.69	0.71	0.70	0.78	0.78	0.77
HVFF ms [52]	0.90	0.94	0.92	0.87	0.91	0.89	0.81	0.81	0.81	0.83	0.76	0.79	0.71	0.73	0.72	0.84	0.66	0.74
GCFE	0.97	0.98	0.97	<b>0.94</b>	<b>0.96</b>	<b>0.95</b>	0.84	<b>0.86</b>	<b>0.85</b>	<b>0.85</b>	<b>0.91</b>	<b>0.88</b>	<b>0.92</b>	<b>0.88</b>	<b>0.90</b>	<b>0.89</b>	<b>0.89</b>	<b>0.89</b>

**Table 3.** Average precision, recall and  $F_1$  scores for all the methods and all the datasets ( $T = 1$ ).

	Synthetic			IDIAP poster			Cocktail party			Coffee break			GDet			Total		
	Prec	Rec	$F_1$	Prec	Rec	$F_1$	Prec	Rec	$F_1$	Prec	Rec	$F_1$	Prec	Rec	$F_1$	Prec	Rec	$F_1$
IRPM [5]	0.53	0.47	0.50	0.71	0.64	0.67	0.28	0.17	0.21	0.27	0.23	0.25	0.59	0.29	0.39	0.46	0.29	0.35
DS [24]	0.68	0.80	0.74	0.79	0.82	0.81	–	–	–	0.40	0.38	0.39	–	–	–	0.60	0.63	0.62
IGD [58]	0.30	0.22	0.25	0.31	0.27	0.29	0.23	0.10	0.13	0.50	0.50	0.50	0.67	0.20	0.31	0.45	0.21	0.27
CTCG [59]	0.78	0.78	0.78	0.83	0.86	0.85	0.31	0.28	0.30	0.46	0.47	0.47	0.51	0.60	0.55	0.49	0.52	0.51
HVFF lin [19]	0.64	0.73	0.68	0.80	0.86	0.83	0.26	0.27	0.27	0.41	0.47	0.44	0.43	0.45	0.44	0.43	0.46	0.44
HVFF ent [51]	0.47	0.52	0.49	0.72	0.74	0.73	0.28	0.30	0.29	0.47	0.52	0.49	0.44	0.45	0.45	0.42	0.44	0.43
HVFF ms [52]	0.72	0.73	0.73	0.73	0.76	0.74	0.30	0.30	0.30	0.40	0.38	0.39	0.44	0.45	0.45	0.44	0.45	0.45
GCFE	<b>0.91</b>	<b>0.91</b>	<b>0.91</b>	<b>0.85</b>	<b>0.87</b>	<b>0.86</b>	<b>0.63</b>	<b>0.65</b>	<b>0.64</b>	<b>0.61</b>	<b>0.64</b>	<b>0.63</b>	<b>0.73</b>	<b>0.68</b>	<b>0.71</b>	<b>0.71</b>	<b>0.70</b>	<b>0.71</b>

**Fig. 4.** Global  $F_1$  score vs. tolerance threshold  $T$ . Between brackets in legend the Global Tolerant Matching score. Dominant Sets (DS) is averaged over 3 datasets only, because of results availability. (Best viewed in colour). (Color figure online)

other kinds of F-formations. Hough-based approaches HVFF X ( $X = lin, ent, ms$ ) have a good modeling of the F-formation, allowing to find any shape, but rely on a greedy optimization procedure. Finally, IRPM approach has a rough modeling of the F-formation. Our approach viceversa has a rich modeling of the F-formation, and a powerful optimization strategy.



**Table 4.** Cocktail party –  $F_1$  score vs. cardinality ( $T = 1$ ).

# groups	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	Avg	Std
	81	82	44	55	147	-	-
IRPM [5]	0.26	0.53	0.74	0.42	0.59	0.51	0.18
IGD [58]	0.06	0.52	0.66	0.73	0.85	0.56	0.30
HVFF lin [19]	0.38	0.76	0.57	0.67	0.94	0.66	0.21
HVFF ent [51]	0.45	0.75	0.69	0.73	0.96	0.71	0.18
HVFF ms [52]	0.49	0.74	0.70	0.71	0.96	0.72	0.17
GCFE	0.59	0.64	0.80	0.85	0.94	<b>0.76</b>	<b>0.14</b>

**Table 5.** GDet –  $F_1$  score vs. cardinality ( $T = 1$ ).

# groups	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	Avg	Std
	197	124	22	35	13	-	-
IRPM [5]	0.40	0.59	0.45	0.42	0.35	0.44	<b>0.09</b>
IGD [58]	0.15	0.52	0.33	0.54	0.83	0.47	0.25
HVFF lin [19]	0.51	0.76	0.03	0.16	0.13	0.32	0.31
HVFF ent [51]	0.57	0.73	0.24	0.23	0.13	0.38	0.26
HVFF ms [52]	0.56	0.78	0.17	0.41	0.67	0.52	0.23
GCFE	0.74	0.87	0.53	0.77	0.88	<b>0.76</b>	0.14

## Cardinality Analysis

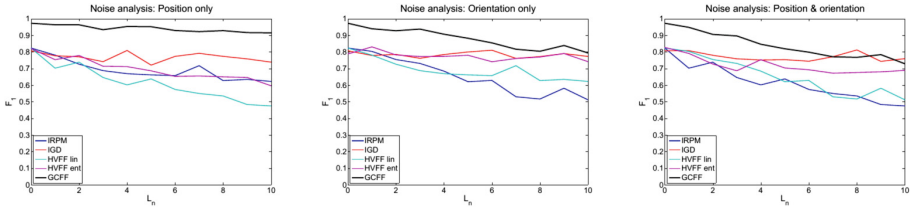
As stated in [52], some methods are shown to work better with some group cardinalities. In this experiment, we systematically check this aspect, evaluating the performance of all the considered methods in individuating groups with a particular number of individuals. Since Synthetic, Coffee Break and IDIAP Poster Session datasets only have groups of cardinality 2 and 3, we only focus on the remaining 2 datasets, which have a more uniform distribution of groups cardinalities. Tables 4 and 5 show  $F_1$  scores for each method and each group cardinality respectively for Cocktail Party and GDet datasets. In both cases the proposed method outperforms the other state of the art methods in terms of higher average  $F_1$  score, with very low standard deviation. In particular, only IRPM gives in GDet dataset results which are more stable than ours, but they are definitely poorer.

## Noise Analysis

In this experiment, we show how the methods behave against different degrees of clutter. For this sake, we consider the Synthetic dataset as starting point and we add to the proxemic state of each individual of each frame some random values based on a known noise distribution. We assume that the noise follows a Gaussian distribution with mean 0, and noise on each dimension (position, orientation) is uncorrelated. For our experiments we used  $\sigma_x = \sigma_y = 20$  cm and  $\sigma_\theta = 0.1$  rad. In our experiments, we consider 11 levels of noise  $L_n = 0, \dots, 10$ , where

$$\begin{cases} x_n(L_n) = x + \text{randsample}(\mathcal{N}(0, L_n * \sigma_x)) \\ y_n(L_n) = y + \text{randsample}(\mathcal{N}(0, L_n * \sigma_y)) \\ \theta_n(L_n) = \theta + \text{randsample}(\mathcal{N}(0, L_n * \sigma_\theta)) \end{cases} \quad (11)$$

In particular, we produce results by adding noise on position only (leaving the orientation at its exact value), on orientation only (leaving the position of each individual at its exact value) and on both position and orientation. Figure 5 shows  $F_1$  scores for each method while increasing the noise level. In this case we can appreciate that with high orientation and combined noise IGD performs comparably or better than GCFE; this is a confirmation of the fact that methods based on Dominant Sets are performing very well when the orientation information is not reliable, as already stated in [51].



**Fig. 5.** Noise analysis.  $F_1$  score vs. noise level on position (left), orientation (centre) and combined (right). (Best viewed in colour). (Color figure online)

## 5 Crowd: Related Work

Abnormal crowd behavior analysis has become an active research topic in recent years, so that several techniques have been proposed to automatically detecting abnormal behavior in crowds. We broadly divide the existing approach in two categories. There are two main categories named (i) *motion-based*, and (ii) *model-based*. In the following, we will give a brief overview of these two categories.

### 5.1 Motion Based Approaches

Motion-based referred to as methods in which motion cues such as optical flow and trajectories are used as a main source information [1, 12, 22, 29, 34, 38–41, 62]. Particularly, initial works on crowd behavior analysis consider crowd as a number of objects interacting in a common space, and the crowd behaviors are inferred according to the motion patterns captured from tracked objects [1, 46, 47]. However, detecting and tracking individuals in a crowd scenes is very challenging, due to imaging condition and density of crowds. In order to tackle this problem, proposed approaches consider crowd as a single entity instead of individual objects. For example, [1] used Finite Time Lyapunov Exponent (FTLE) field to detect and localize anomaly in a frame level from crowd videos by measuring flow instability in the segmented regions of high density crowd flows. However, this approach is limited to the structural scene where the boundary of crowd flow is well determined. To tackle this problem, a chaotic invariant approach is proposed for structural and nonstructural crowd scenes [61]. Particularly, the chaotic dynamics of crowd are computed from the crowd flow trajectory. The main limitation of this method is its computational complexity, since it use all the previous frames. Moreover, there exist also a significant number of research focus on identifying specific type of abnormality (e.g., violence, panic). The first work for detecting violence in videos was proposed in [20], which mainly focused on two person fight episodes and employed motion trajectory information of individual limbs for fight modeling and classification. This approach required limbs segmentation and tracking, which are very challenging tasks in a presence of occlusion and clutters. In [22] authors exploit the statistics magnitude of optical flow varying along the time for detecting violence behavior from dense crowded scenes. However, the effectiveness of this approach is also limited on

highly dense crowd. In [40] authors quantified abnormality exploiting the statistic of motion estimated from magnitude and orientation of tracklet (short portion of trajectory) in terms of two dimensional histogram. Although, their method show promising performance on a wider ranger of crowd anomaly, choice of efficient number of bins across the scenes is under the question. In [38] authors used substantial derivative, a well-known concept in fluid dynamic, for extracting meaningful motion patterns to identify violence behaviors from video sequences. Moreover, authors demonstrate the effectiveness of the method on panic scenes. In the following we will provide brief details on [20, 22, 38] related to the motion based approaches.

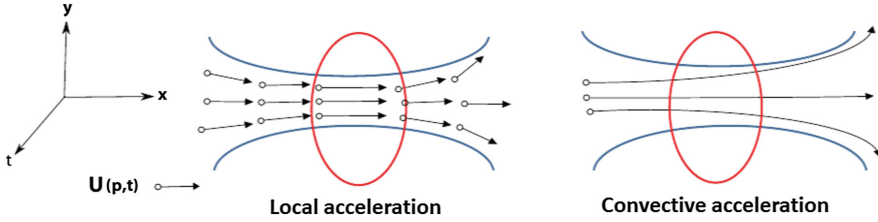
**Acceleration Measure Vector (Jerk).** This approach can be considered as one of the first attempt in detecting human violence in video [20]. To detect violent behavior authors rely on motion trajectory information and on orientation information of a persons limbs. Then an Acceleration Measure Vector (AMV) composed of direction and magnitude of motion and jerk is defined to be the temporal derivative of AMV.

**Violence Flows (ViF).** It is specifically designed for classifying violent behaviors in crowds from video outbreaks [22]. The main strength of ViF lies in its ability in encoding meaningful temporal motion patterns. Initially, the magnitude of motions are estimated. Next, for each video frame a binary map is created by discarding motion values less than a predefined threshold. Then, final motion map is created by computing average along all video frames in a query video.

**Substantial Derivative.** Substantial Derivative (SD) is an important concept in fluid mechanics which describes the change of fluid elements by physical properties such as temperature, density, and velocity components of flowing fluid along its trajectory [4]. Unlike aforementioned approaches that only use temporal motion patterns as a main source of information [22, 36, 40], it has a capability of capturing spatial and temporal information of motion changes in a single framework [4]. Specifically, SD consists of two terms of accelerations, namely *local* and *convective* accelerations. The local acceleration captures the change rate of velocity of a certain particle with respect to time and vanishes if the flow is steady. On the other hands, the convective acceleration captures the change of velocity flow in the spatial space, therefore, it increases when particles move through the region of spatially varying velocity field in the temporal domain (see Fig. 6 for an example). Therefore, the local acceleration, which is the derivative of velocity with respect to the time, captures the instant change of flow. Whilst, the convective acceleration captures the spatial evolution of a particle moving along its trajectory. This is, in particular, useful in the abnormal crowd scenarios. Considering violence as an example of abnormal scene, where individual shows aggressive behavior toward other member of crowd, his/her motion is subject to sudden change of in velocity field with respect to the time

(local acceleration). Moreover, at the same time, group of individual interact in a very different way compare with violence (or panic scenes), which lead to the significant difference in the motion trajectories of individuals compared with normal scenes in spatial domain (convective acceleration). For video representation random patch sampling with Bag-of-Word Paradigm applied from each force independently. Finally, the local and convective forces are concatenated to shape a single descriptor named *total force*.

Tables 6 and 7 provide some detail regarding the aforementioned motion based methods, including their general formula and force estimation from video sequences.



**Fig. 6.** An example of local and convective accelerations. The local acceleration measure instantaneous rate of change of each particle, while convective acceleration measures the rate of change of the particle moving along its trajectory. Red region indicates the particles are accelerated as it converge due to the structural change of the environment. (Color figure online)

**Table 6.** Summary of motion and model based forces, their general formula and estimation methods from video sequences.

Methods	Forces	General formula	Estimation from videos
Behavioral heuristics [49]	$H_1$	$\frac{dv_i}{dt} = \frac{v_i^{des} - v_i(t)}{\tau}$	$\frac{dv_i}{dt} \approx \mathbf{a}_i^{t+1} = \mathbf{O}_i^{t+1} - \mathbf{O}_i^t$
	$F_{ij}^{bc}(H_2)$	$F_{ij}^{bc} = n_{ji} \cdot \mathbf{g}_i(j)$	$F_i^{bc} = \frac{\sum_j \mathbf{a}_j \cdot \mathbf{g}_i(j)}{\sum_j \mathbf{g}_i(j)}$
	$F_{ij}^{agg}(H_3)$	$F_{ij}^{agg} = n_{ji} \cdot \frac{(1 - \frac{v_i \cdot v_j}{\ v_j\  \cdot \ v_i\ })}{2} \cdot \mathbf{f}_i(j)$	$F_i^{agg} = \sum_j (n_{ji} \cdot \mathbf{w}_{ij} \cdot \mathbf{f}_{\mathbf{O}_i}^\alpha(j))$
SFM [36]	$F_{int}$	$F_{int} = \frac{1}{\tau} (v_i^q - v_i) - \frac{dv_i}{dt}$	$v_i = \langle \mathbf{O}_i \rangle$ $v_i^q = (1 - p_i)\mathbf{O}_i + p_i \langle \mathbf{O}_i \rangle$
Substantial derivative [38]	$F^L$	$F^L = m_i \frac{dv_i}{dt}$	$\frac{dv_i}{dt} \approx \mathbf{a}_i^{t+1} = \mathbf{O}_i^{t+1} - \mathbf{O}_i^t$
	$F^{Cv}$	$F^{Cv} = m_i \sqrt{a_x^2 + a_y^2}$	$\mathbf{a}_x = \left( \frac{\partial \langle \mathbf{O}_x \rangle}{\partial x} + \frac{\partial \langle \mathbf{O}_y \rangle}{\partial y} \right) \cdot \langle \mathbf{O}_x \rangle$ $\mathbf{a}_y = \left( \frac{\partial \langle \mathbf{O}_x \rangle}{\partial x} + \frac{\partial \langle \mathbf{O}_y \rangle}{\partial y} \right) \cdot \langle \mathbf{O}_y \rangle$
	$F^T$	$F^T = F^{Cv} + F^L$	$F^T = F^{Cv}   F^L$
Jerk [20]	$J$	$J_i = \frac{d\mathbf{a}_i}{dt}$	$\frac{d\mathbf{a}_i}{dt} \approx \mathbf{J}_i = \mathbf{a}_i^{t+1} - \mathbf{a}_i^t$

## 5.2 Model Based Approaches

Model based approaches can be broadly categorized into two categories; (i) The Social Force Model (SFM), (ii) Behavioral heuristic models.

**Table 7.** Summary of notations.

Notations	Description formula
$\mathbf{v}_i/\mathbf{O}_i^t$	Velocity of particles/individuals $i$
$\langle \mathbf{O}_i \rangle$	Average velocity of particles/individuals $i$
$g(i)$	Body compression subject to the individual $i$ and neighboring individuals
$W_{i,j}$	Aggression factor between the individual $i$ and his/her opponent $j$
$\theta_{i,j}$	Orientation between individual $i$ and his/her opponents $j$
$d_{i,j}$	Distance between individual $i$ and $j$
$R$	Personal space around each examined individual
$p_i$	Panic factor
$f(i)$	Model the angle of view of individual $i$ individual
$\mathbf{a}_x$	Spatial acceleration in x direction
$\mathbf{a}_y$	Spatial acceleration in y direction
$\mathbf{F}_{ij}^{bc}$	Body contact individual $i$ with its neighboring pedestrians $j$
$\mathbf{F}_{ij}^{agg}$	Aggression force pedestrian $i$ derived toward its opponent $j$
$\mathbf{F}^L$	Local force
$\mathbf{F}^{Cv}$	Convective force
$\mathbf{F}^T$	Total force

**The Social Force Model.** The SFM originally introduced by Helbing and Molnar [23], is the historical seminal method for modeling crowd behaviors according to a set of predefined physical rules. More specifically, the SFM aimed at representing the interaction force among individuals in crowded scenes using a set of *repulsive* and *attractive* forces, which was shown to be a significant feature for analyzing crowd behaviors. Motivated by the success of SFM to reproduce crowd moving patterns, fifteen years later Mehran et al. [36] adopted the SFM and particle advection scheme to compute for detecting and localizing abnormal behavior in crowd videos. To this purpose, they considered the entire crowd as a set of moving particles whose interaction force was computed using SFM. Then, they mapped the interaction force into the image plane to obtain the force flow of each particle within frame of videos. This force map was used as the basis for extracting features which, along with the random spatial-temporal path sampling and BOW strategy, was used to assign either normal or abnormal label to each frame. Moreover, the force map was also employed to localize the anomaly region in the detected abnormal frames.

### 5.3 Behavioral Heuristics

Behavioral heuristic approach can be considered as a new emergent approach to disclose complex crowd dynamics [42, 43], compare to SFM [23], where single Newtonian equation used to explain crowd behaviors, it exploits the use

of physics equations inspired from simple, yet effective behavioral heuristics to describe the crowd behaviors. This is, in particular useful since it is capable of capturing wider range of crowd complexities. To this end, Mohammadi et al. [49], proposed a heuristic based method inspired from [42, 43], and successfully applied for violence detection in crowds [49]. Their method established based on three heuristic rules:

- H1: *An individual chooses the direction that allows the most direct path to a destination point, adopting his/her moving regarding the presence of obstacles.*  
 H2: *In crowd situations, the movement of an individual is influenced by his/her physical body contacts with surrounding persons.*  
 H3: *In violent scenes, an individual mainly moves towards his/her opponents to display violent actions.*

Specifically, (H1) describe individual’s internal motivation towards a goal avoiding obstacles or other individuals. While, the second heuristic rule (H2), states that individual movements are subject to the unintentional physical body contacts with his/her surrounding individuals. The third heuristic rule (H3) defines behavioral patterns within violent scenes, where there are two or more parties (e.g., police and rioters) fighting and showing violent behaviors to each other.

Then, each heuristic is formulated with physics equation(see Tables 6 and 7 for details). Next, each force is computed, independently, from video sequences. Finally, random patch sampling along with Bag-of-Word paradigm used for force representation, and all the forces are concatenated to construct the final descriptor named Visual Information Processing Signature (VIPS).

## 6 Crowds: Experiments

We extensively evaluated proposed approach on five benchmarks, consists of two standard datasets; Violence in Crowds [22], and Behave [8] along with two video sequences collected from web source (i.e., [www.YouTube.com](http://www.YouTube.com)) which we named Panic1 and Panic2. Moreover, we assembled a new dataset, named “Violent-Cross” whose videos gathered from Violence in Crowds and CUHK [54] datasets, to show the ability of proposed approaches in cross scene recognition. Specifically, it includes 300 videos, equally divided into three classes (100 videos for each class). Figure 7 shows few samples of the benchmarks. For the Violence in Crowds we used the standard training/testing split, released along with dataset. While for Behave, panic1 and panic2, we divide the video frames into block of 10 frames with one frame overlapping among the blocks, then we labeled them into normal and abnormal blocks. For Violence-Cross, we equally divide each class into a test set of 150 videos (50 video sequences for each class) and the rest for testing. For feature representation we used Bag-of-Word Paradigm. Particularly, first we perform random 3D patch sampling from the computed force maps, then Bag-of-Word Paradigm, with fixed number of cluster centers to generate codebook. Finally, the resulting histogram of visual words are fed into the classifiers. As a

choice of classifier, for Violence in Crowds and Violence-Cross, we used SVM, since we had access to negative data at the training time. While, for other datasets, we used Latent Dirichlet Allocation [7] in short LDA since we assumed that we do not have access to negative data at the training time.



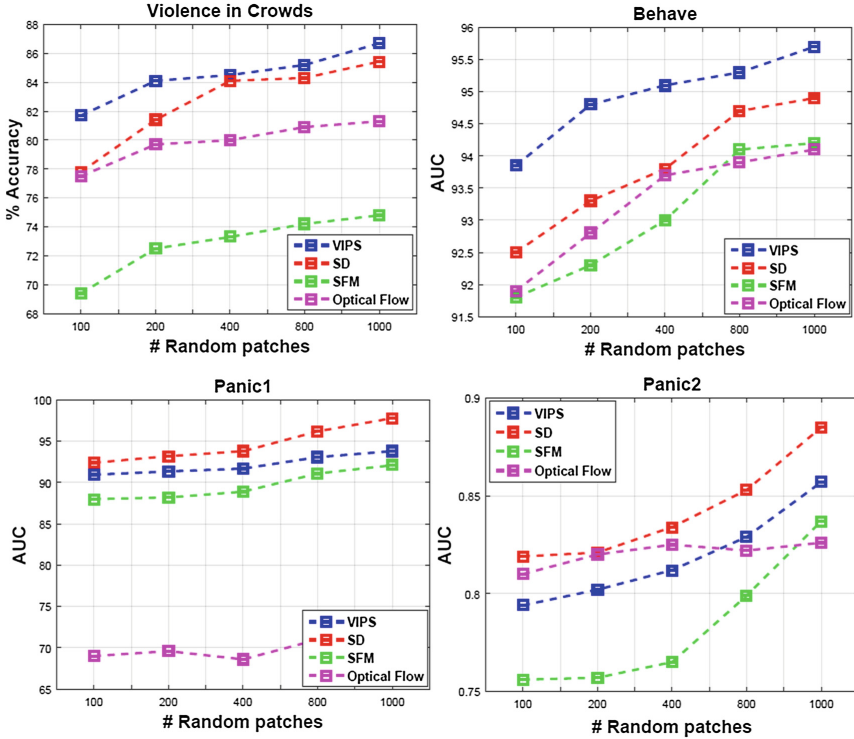
**Fig. 7.** Sample frames from the evaluated datasets. For first four columns; first row shows the normal crowd behaviors, while for the second row shows the abnormal behavior, specifically, violent and panic behaviors. For Violence-Cross datasets; samples from the three classes of “normal”, “crossing” and “violent” behaviors.

### 6.1 Effect of the Random Patches

Here, we examine the effect of varying number of random patches in the range of  $P \in \{100, 200, 400, 800, 1000\}$  on various datasets by fixing number of cluster center to 500. Figure 8 shows the effect of varying number of patches on optical flow, SFM [36], SD ( $F^L|F^{Cv}$ ) [38], and VIPS [49]. As it is visible, all the descriptors show improvement in performance by increasing the number of random patches, however, we did not observe any significant improvement after reaching 1000 random patches. Moreover, we observed that VIPS shows very promising performance compare with other methods on Violence in Crowds and Behave datasets. While, SD shows its superiority in performance on panic scenes. This is understandable, since VIPS specifically designed for violence detection and aggression force plays an important role in capturing violence behavioral patterns. While, experimental results show that in panic scenarios (Panic1 and Panic2 datasets) combination of temporal and structure information in SD descriptor offers more discriminative features compare with other proposed methods.

### 6.2 Comparison with the State of the Art

We compared various motion and model based methods for the evaluation purpose. Particularly, for the motion based approaches, we select ViF [22], Jerk [20], total force inspired from SD [38], and for model based approach we select the SFM [36], and recently proposed heuristic based model in [49].



**Fig. 8.** Effect of varying number of random patches on the effectiveness of VIPS [49], SD [38], SFM [36] and optical flow on Violence in Crowds, Behave, Panic1 and Panic2 datasets.

Table 8 show the comparison of motion and model based approaches. It is visible that among the compared methods heuristic based method outperform other competitors especially in Violence in Crowds dataset, this support the psychology studies [42,43], and highlight the strength of heuristic models in capturing wider range of crowd complexities in violence scenes which results in better performance. It is also visible that SD [38] can be considered as a close competitor to the heuristic method. Indeed, it outperforms the heuristic-based method with a high margin in Panic datasets. This is understandable since, VIPS [49] specifically designed for violence detection. In addition, it has a consist of spatial information (Local and Convective Forces, respectively), which make it capable in covering wider range of crowd complexity compared with other methods.

Moreover, we evaluate robustness of our descriptors to distinguish between acts of violence from crossing behaviors, which is a most similar approach to the act of violence. In particular, as a motion based descriptor, we select ViF [22], which is specifically design for violence detection, along with SFM [36], which is considered as one of the most well-known descriptor to detect abnormality



**Table 8.** Comparison of model and motion based approaches.

Descriptors	Violence in Crowds	Behave	Panic1	Panic2
ViF [22]	81.3	93.4	0.716	82.6
AMV (Jerk) [20]	74.8	94.2	90.9	83.6
SD ( $F^L F^{Cv}$ ) [38]	85.4	94.8	98.5	88.4
SFM [36]	74.5	94.23	91.3	84.3
VIPS [49]	86.61	95.73	94.5	85.9

ViF				SFM				VIPS			
Class 1	0.94	0.02	0.04	Class 1	0.86	0.12	0.02	Class 1	0.96	0.04	0
Class 2	0.36	0.30	0.34	Class 2	0.26	0.72	0.02	Class 2	0.06	0.94	0
Class 3	0.48	0.04	0.48	Class 3	0.37	0.12	0.51	Class 3	0.1	0.02	0.88
	Class 1	Class 2	Class 3		Class 1	Class 2	Class 3		Class 1	Class 2	Class 3

**Fig. 9.** Average accuracy on Violent-Cross dataset. Class1, Class2, and Class3 are referred to as violent, cross walk, and normal behaviors, respectively. ViF [22] with 57% overall accuracy; SFM [36] with 69% overall accuracy, and VIPS with 92% overall accuracy.

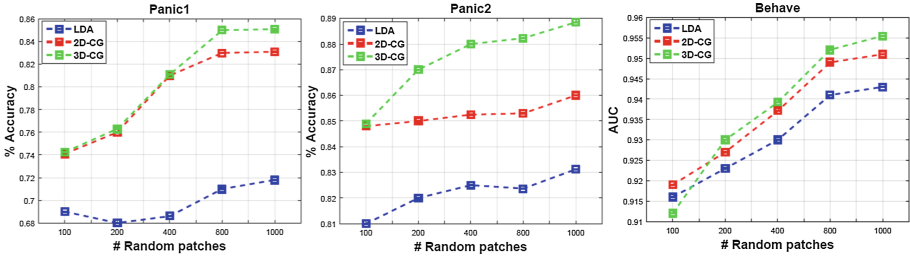
in crowds. Figure 9 shows the confusion matrices of two state-of-the-art methods and elements of the proposed method. We observe that ViF shows a good performance on detecting acts of violence compared to the SFM, however, its overall accuracy is low since it is incapable of distinguishing violent from normal and crossing behaviors. On the other hand, we observe that VIPS shows very promising performance compared with other approaches.

**Table 9.** Average AUCs on Panic1, Panic2 and Behave datasets sequences with 1000 random patches and 500 cluster centers.

Generative model	Datasets		
	Panic1	Panic2	BEHAVE
LDA [7]	0.718	0.826	0.943
2D-CG [26]	0.833	0.86	0.951
3D-CG [37]	0.851	0.865	0.958

### 6.3 Choice of Generative Model

In general, there exist two methods for evaluating the effectiveness of descriptors for violence detection; Discriminative (e.g., SVM) and Generative (e.g., LDA [7], Counting Grids [26]) methods. Although, discriminative approaches are considered as more powerful approach for detecting abnormalities, it requires negative



**Fig. 10.** Comparison effect of random patches of average AUCs (y-axis) on Panic1, Panic2, and Behave datasets, using LDA, 2D-CG and 3D-CGs.

examples at the training time. While, the definition of abnormality is still context dependent, therefore gathering huge amount of negative data is very challenging problem. To this end, one common practice to face this problem, is to learn what normality is and then abnormality considered motion patterns which deviated from learned distribution of normal behavior [36, 37]. However, apart from effectiveness of descriptors in capturing crowd complexities, capability of generative models on modeling the distribution of normal data plays an important roles in detecting abnormal behaviors. To this end, we show comparison on Generative models. Particularly, we select the optical flow as a baseline, and we compared LDA with two dimensional Counting Grid (2D-CG) [26] and three dimensional Counting Grid (3D-CGs) [37] on Panic1, Panic2 and Behave datasets. Table 9 shows that 3D-CG outperform two competitors. This is understandable, since 3D-CG is able to capture spatial-temporal relationship among the bags in the feature space, while CG is only consider spatial information and LDA totally ignore intra relationship among the bags (Fig. 10).

## 7 Conclusions

In this chapter, we provide a broad overview along with extensive experiments of most recent state of the art methods in group detection and crowd behaviors understanding. The experimental results demonstrate that pure computer vision techniques may not be sufficient to uncover wide range of group and crowds' behaviors/dynamics, and sociological-inspired methodologies outperform other state-of-art approaches. To this end, we believe that proposing an universal approach for human behavior understanding in group and crowd levels is still can be considered as open problems, and further investigations are required to introduce new methodologies in computer vision community.

## References

1. Ali, S., Shah, M.: A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–6. IEEE (2007)

2. Ali, S., Shah, M.: Floor fields for tracking in high density crowd scenes. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5303, pp. 1–14. Springer, Heidelberg (2008). doi:[10.1007/978-3-540-88688-4\\_1](https://doi.org/10.1007/978-3-540-88688-4_1)
3. Ba, S.O., Odobez, J.M.: Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* **33**(1), 101–116 (2011)
4. Batchelor, G.K.: *An Introduction to Fluid Dynamics*. Cambridge University Press, Cambridge (2000)
5. Bazzani, L., Cristani, M., Tosato, D., Farenzena, M., Paggetti, G., Menegaz, G., Murino, V.: Social interactions by visual focus of attention in a three-dimensional environment. *Expert Syst.* **30**(2), 115–127 (2013)
6. Benfold, B., Reid, I.: Stable multi-target tracking in real-time surveillance video. In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3457–3464 (2011)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003)
8. Blunsden, S., Fisher, R.: The BEHAVE video dataset: ground truthed video for multi-person behavior classification. *Ann. BMVA* **2010**(4), 1–12 (2010)
9. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* **23**(11), 1222–1239 (2001)
10. Bozdogan, H.: Model selection and Akaike’s Information Criterion (AIC): the general theory and its analytical extensions. *Psychometrika* **52**(3), 345–370 (1987)
11. Campbell, N.D.F., Vogiatzis, G., Hernández, C., Cipolla, R.: Automatic 3D object segmentation in multiple views using volumetric graph-cuts. *Image Vis. Comput.* **28**(1), 14–25 (2010)
12. Cao, T., Wu, X., Guo, J., Yu, S., Xu, Y.: Abnormal crowd motion analysis. *ROBIO* **9**, 1709–1714 (2009)
13. Chan, A.B., Liang, Z.S.J., Vasconcelos, N.: Privacy preserving crowd monitoring: counting people without people models or tracking. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008*, pp. 1–7. IEEE (2008)
14. Chen, C., Odobez, J.M.: We are not contortionists: coupled adaptive learning for head and body orientation estimation in surveillance video. In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1544–1551 (2012)
15. Ciolek, T.M.: The proxemics lexicon: a first approximation. *J. Nonverbal Behav.* **8**(1), 55–79 (1983)
16. Ciolek, T.M., Kendon, A.: Environment and the spatial arrangement of conversational encounters. *Sociol. Inq.* **50**(3–4), 237–271 (1980)
17. Conte, D., Foggia, P., Percannella, G., Tufano, F., Vento, M.: A method for counting people in crowded scenes. In: *2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 225–232. IEEE (2010)
18. Cook, M.: Experiments on orientation and proxemics. *Hum. Relat.* **23**(1), 61–76 (1970)
19. Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Del Bue, A., Menegaz, G., Murino, V.: Social interaction discovery by statistical analysis of formations. In: *British Machine Vision Conference (BMVC)*, pp. 23.1–23.12 (2011)
20. Datta, A., Shah, M., Lobo, N.D.V.: Person-on-person violence detection in video data. In: *Proceedings 16th International Conference on Pattern Recognition 2002*, vol. 1, pp. 433–438. IEEE (2002)

21. Gong, S., Loy, C.C., Xiang, T.: Security and surveillance. In: Moeslund, T.B., Hilton, A., Krüger, V., Sigal, L. (eds.) *Visual Analysis of Humans*. Springer, Heidelberg (2011). doi:[10.1007/978-0-85729-997-0\\_23](https://doi.org/10.1007/978-0-85729-997-0_23)
22. Hassner, T., Itcher, Y., Kliper-Gross, O.: Violent flows: real-time detection of violent crowd behavior. In: *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–6. IEEE (2012)
23. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. *Phys. Rev. E* **51**(5), 4282 (1995)
24. Hung, H., Kröse, B.: Detecting F-formations as dominant sets. In: *International Conference on Multimodal Interfaces (ICMI)*, pp. 231–238 (2011)
25. Jiang, F., Wu, Y., Katsaggelos, A.K.: Detecting contextual anomalies of crowd motion in surveillance video. In: *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 1117–1120. IEEE (2009)
26. Jojic, N., Perina, A.: Multidimensional counting grids: inferring word order from disordered bags of words. *arXiv preprint* (2012). [arXiv:1202.3752](https://arxiv.org/abs/1202.3752)
27. Kendon, A.: *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge University Press, Cambridge (1990)
28. Kok, V.J., Lim, M.K., Chan, C.S.: Crowd behavior analysis: a review where physics meets biology. *Neurocomputing* **177**, 342–362 (2015)
29. Kratz, L., Nishino, K.: Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 1446–1453. IEEE (2009)
30. Ladický, L., Russell, C., Kohli, P., Torr, P.H.S.: Inference methods for CRFs with co-occurrence statistics. *Int. J. Comput. Vis.* **103**(2), 213–225 (2013)
31. Lanz, O.: Approximate Bayesian multibody tracking. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* **28**(9), 1436–1449 (2006)
32. Lanz, O., Brunelli, R.: Joint Bayesian tracking of head location and pose from low-resolution video. In: Stiefelhagen, R., Bowers, R., Fiscus, J. (eds.) *CLEAR/RT-2007*. LNCS, vol. 4625, pp. 287–296. Springer, Heidelberg (2008). doi:[10.1007/978-3-540-68585-2\\_27](https://doi.org/10.1007/978-3-540-68585-2_27)
33. Li, T., Chang, H., Wang, M., Ni, B., Hong, R., Yan, S.: Crowded scene analysis: a survey. *IEEE Trans. Circuits Syst. Video Technol.* **25**(3), 367–386 (2015)
34. Liu, Y., Li, X., Jia, L.: Abnormal crowd behavior detection based on optical flow and dynamic threshold. In: *2014 11th World Congress on Intelligent Control and Automation (WCICA)*, pp. 2902–2906. IEEE (2014)
35. Lombaert, H., Sun, Y., Grady, L., Xu, C.: A multilevel banded graph cuts method for fast image segmentation. In: *International Conference on Computer Vision (ICCV)*, pp. 259–265 (2005)
36. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, pp. 935–942. IEEE (2009)
37. Mohammadi, S., Kiani, H., Perina, A., Murino, V.: A comparison of crowd comotion measures from generative models. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 49–55 (2015)
38. Mohammadi, S., Kiani, H., Perina, A., Murino, V.: Violence detection in crowded scenes using substantial derivative. In: *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6. IEEE (2015)
39. Mousavi, H., Galoogahi, H.K., Perina, A., Murino, V.: Detecting abnormal behavioral patterns in crowd scenarios. In: Esposito, A., Jain, L.C. (eds.) *Toward Robotic Socially Believable Behaving Systems - Volume II*. ISRL, vol. 106, pp. 185–205. Springer, Cham (2016). doi:[10.1007/978-3-319-31053-4\\_11](https://doi.org/10.1007/978-3-319-31053-4_11)

40. Mousavi, H., Mohammadi, S., Perina, A., Chellali, R., Murino, V.: Analyzing tracklets for the detection of abnormal crowd behavior. In: 2015 IEEE Winter Conference on Applications of Computer Vision, pp. 148–155. IEEE (2015)
41. Mousavi, H., Nabi, M., Kiani, H., Perina, A., Murino, V.: Crowd motion monitoring using tracklet-based commotion measure. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 2354–2358. IEEE (2015)
42. Moussaïd, M., Helbing, D., Theraulaz, G.: How simple rules determine pedestrian behavior and crowd disasters. *Proc. Nat. Acad. Sci.* **108**(17), 6884–6888 (2011)
43. Moussaïd, M., Nelson, J.D.: Simple heuristics and the modelling of crowd behaviours. In: Weidmann, U., Kirsch, U., Schreckenberg, M. (eds.) *Pedestrian and Evacuation Dynamics 2012*, pp. 75–90. Springer, Cham (2014). doi:[10.1007/978-3-319-02447-9\\_5](https://doi.org/10.1007/978-3-319-02447-9_5)
44. Pavan, M., Pelillo, M.: Dominant sets and pairwise clustering. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* **29**(1), 167–172 (2007)
45. Poppe, R.: A survey on vision-based human action recognition. *Image Vis. Comput.* **28**(6), 976–990 (2010)
46. Rabaud, V., Belongie, S.: Counting crowded moving objects. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), vol. 1, pp. 705–711. IEEE (2006)
47. Rittscher, J., Tu, P.H., Krahnstoeber, N.: Simultaneous estimation of segmentation and shape. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 2, pp. 486–493. IEEE (2005)
48. Rodriguez, M., Ali, S., Kanade, T.: Tracking in unstructured crowded scenes. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1389–1396. IEEE (2009)
49. Mohammadi, S., Perina, A., Kiani, H., Murino, V.: Angry crowds: detecting violent events in videos. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016*. LNCS, vol. 9911, pp. 3–18. Springer, Cham (2016). doi:[10.1007/978-3-319-46478-7\\_1](https://doi.org/10.1007/978-3-319-46478-7_1)
50. Saleh, S.A.M., Suandi, S.A., Ibrahim, H.: Recent survey on crowd density estimation and counting for visual surveillance. *Eng. Appl. Artif. Intell.* **41**, 103–114 (2015)
51. Setti, F., Hung, H., Cristani, M.: Group detection in still images by f-formation modeling: a comparative study. In: *International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pp. 1–4 (2013)
52. Setti, F., Lanz, O., Ferrario, R., Murino, V., Cristani, M.: Multi-scale F-formation discovery for group detection. In: *IEEE International Conference on Image Processing (ICIP)* (2013)
53. Setti, F., Russell, C., Bassetti, C., Cristani, M.: F-formation detection: individuating free-standing conversational groups in images. *PloS ONE* **10**(5), e0123783 (2015)
54. Shao, J., Loy, C.C., Wang, X.: Scene-independent group profiling in crowd. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2219–2226 (2014)
55. 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 Trans. Pattern Anal. Mach. Intell. (PAMI)* **30**(7), 1212–1229 (2008)
56. Tang, S., Andriluka, M., Milan, A., Schindler, K., Roth, S., Schiele, B.: Learning people detectors for tracking in crowded scenes. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1049–1056 (2013)

57. Tosato, D., Spera, M., Cristani, M., Murino, V.: Characterizing humans on Riemannian manifolds. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* **35**(8), 1972–1984 (2013)
58. Tran, K.N., Bedagkar-Gala, A., Kakadiaris, I.A., Shah, S.K.: Social cues in group formation and local interactions for collective activity analysis. In: *International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 1, pp. 539–548 (2013)
59. Vascon, S., Mequanint, E.Z., Cristani, M., Hung, H., Pelillo, M., Murino, V.: A game-theoretic probabilistic approach for detecting conversational groups. In: *Creemers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS*, vol. 9007, pp. 658–675. Springer, Cham (2015). doi:[10.1007/978-3-319-16814-2\\_43](https://doi.org/10.1007/978-3-319-16814-2_43)
60. Vascon, S., Mequanint, E.Z., Cristani, M., Hung, H., Pelillo, M., Murino, V.: Detecting conversational groups in images and sequences: a robust game-theoretic approach. *Comput. Vis. Image Underst.* **143**, 11–24 (2016)
61. Wu, S., Moore, B.E., Shah, M.: Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes. In: *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2054–2060. IEEE (2010)
62. Xu, L., Gong, C., Yang, J., Wu, Q., Yao, L.: Violent video detection based on MoSIFT feature and sparse coding. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3538–3542. IEEE (2014)
63. Xu, N., Ahuja, N., Bansal, R.: Object segmentation using graph cuts based active contours. *Comput. Vis. Image Underst.* **107**(3), 210–224 (2007)
64. Zhan, B., Monekosso, D.N., Remagnino, P., Velastin, S.A., Xu, L.Q.: Crowd analysis: a survey. *Mach. Vis. Appl.* **19**(5–6), 345–357 (2008)