Social interaction discovery by statistical analysis of F-formations

Marco Cristani[1,2]
marco.cristani@univr.it
Loris Bazzani[1]
loris.bazzani@univr.it
Giulia Paggetti[1]
giulia.paggetti@univr.it
Andrea Fossati[3]
fossati@vision.ee.ethz.ch
Diego Tosato[1]
diego.tosato@univr.it
Alessio Del Bue[2]
alessio.delbue@iit.it
Gloria Menegaz[1]
gloria.menegaz@univr.it
Vittorio Murino[1,2]
vittorio.murino@iit.it

[1] Dipartimento di Informatica
University of Verona
Italy

[2] Istituto Italiano di Tecnologia – IIT
Via Morego, 30
16163 Genova, Italy

[3] ETH Zürich
Sternwartstrasse 7
CH - 8092 Zürich, Switzerland

Abstract

We present a novel approach for detecting social interactions in a crowded scene by employing solely visual cues. The detection of social interactions in unconstrained scenarios is a valuable and important task, especially for surveillance purposes. Our proposal is inspired by the social signaling literature, and in particular it considers the sociological notion of F-formation. An F-formation is a set of possible configurations in space that people may assume while participating in a social interaction. Our system takes as input the positions of the people in a scene and their (head) orientations; then, employing a voting strategy based on the Hough transform, it recognizes F-formations and the individuals associated with them. Experiments on simulations and real data promote our idea.

1 Introduction

Detecting human interactions represents one of the most intriguing frontiers in the automated surveillance since more than a decade [1, 25]. Very recently, sociologic reasoning has been incorporated into video-surveillance algorithms, thanks to the social signalling studies; this trend has rapidly grown, as witnessed in the literature [9, 19, 26, 27, 29, 31].

This paper goes in this direction as it attempts to discover social interactions using statistical analysis of spatial-orientational arrangements that have a sociological relevance. As

BMVC 2011 http://dx.doi.org/10.5244/C.25.23

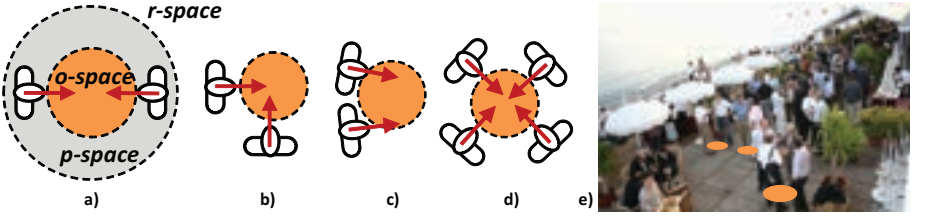


Figure 1: F-formations: a-d) The component spaces of an F-formation: vis-a-vis, L, side-by-side, and circular F-formations, respectively. O-spaces are drawn in orange. e) Cocktail-party scene where some o-spaces are superimposed in orange.

social interactions we intend *the acts, actions, or practices of two or more people mutually oriented towards each other's selves, that is, any behavior that tries to affect or take account of each other's subjective experiences or intentions* [30]. For instance, talking is the most common kind of social interaction. Working together, playing chess, eating at a table and offering a cup of water are social interactions too. In general, any dynamic sequence of social actions between individuals (or groups) that modify their actions and reactions by their interaction partner(s) are social interactions.

In our case, we analyze quasi-stationary people in an unconstrained scenario identifying those subjects engaged in a face-to-face interaction, i.e., a scene monitored by a single camera where a variable amount of people (10-20) is present. We import into the analysis the sociological concept of F-formation as defined by Adam Kendon in the late '70s [14, 15, 16, 17], commonly adopted in the sociological literature. Simply speaking, F-formations are spatial patterns maintained during social interactions by two or more people. Quoting Kendon, "*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*.". In practice, an F-formation is the proper organization of three social spaces: o-space, p-space and r-space (see Fig. 1a-d).

The o-space is a convex empty space surrounded by the people involved in a social interaction, where every participant looks inward into it, and no external people is allowed in this region. This is the most important part of an F-formation. The p-space is a narrow stripe that surrounds the o-space, and that contains the bodies of the talking people, while the r-space is the area beyond the p-space.

There can be different F-formations as visible in Fig. 1a-d. In the case of two participants, typical F-formation arrangements are vis-a-vis, L-shape, and side-by-side. When there are more than three participants, a circular formation is typically formed [18].

Our approach aims at detecting the o-space, taking as input a calibrated scenario, in which the position of the people and their head's orientations have been estimated. In particular, we design an F-formation recognizer which is the main contribution of the work. This algorithm is based on a Hough-voting strategy, which lies between an implicit shape model [21], where weighted local features vote for a location in the image plane, and a mere generalized Hough procedure where the local features have not to be in a fixed number as in the implicit shape model. This approach provides the estimation of the o-spaces, so as of the identity of the people that form them, thus individuating people which are socially interacting. In such regard, our approach is the first to use F-formations detection in order to discover social interactions solely from visual cues.

Our approach has been tested on about a hundred of simulated scenarios, and two real annotated datasets, one of which is novel. In these last two cases tens of individuals were

captured while they were enjoying coffee breaks, in indoor and outdoor environment, giving rise to heterogeneous real crowded scenarios. Our approach obtains convincing results, that are reported in a comparative way, quoting the unique (to the best of our knowledge) previous work dealing with the same topic.

The rest of the paper is organized as follows. In Sec.2, a review of the literature concerning the interaction modelling in surveillance settings is given. The proposed approach is detailed in Sec. 3, and the experiments are reported in Sec. 4. Finally, Sec. 5 concludes the paper with remarks and a discussion on the several possible future developments.

2 State of the Art

A dated but interesting review on methods that consider human interactions is presented in [1], that focuses especially on motion cues. Pioneering studies on interactions focus on two-agent behaviors, employing statistical learning [25], a mix between syntactical and statistical pattern recognition paradigms [12], or Action-Reaction Learning [13]. Interactions among a larger number of people are usually modeled in meeting scenarios or smart rooms, exploiting a large number of heterogeneous sensors, thus solving many problems of occlusions and low image quality. In this case, many subtle social interactions can be observed and modeled, mostly by encoding turn-taking mechanisms. The interested reader may refer to [8] for a comprehensive review. Moving to unconstrained scenarios, as those typical of the video-surveillance field, the spectra of the activities modeled becomes narrower. In [11] a Semi Markov framework captures simple events (as running, approaching, etc.), where interaction is modeled by logic operators that assembly together simple events (performed by a single person) into multi-thread events. More recently, in [6, 24], group activities are encoded with three types of localized causalities, namely self-causality, pair-causality, and group-causality, which characterize the local interaction/reasoning relations within, between, and among motion trajectories of different humans, respectively. In [22], group interactions with a varying number of subjects are investigated, employing an asynchronous hidden Markov model as a hierarchical activity model. They distinguish symmetric (like *i* talks with *j*) and asymmetric dynamics activities (like *i* follows *j*). A discriminative approach is proposed in [19], in which two kinds of interactions are introduced. The first, group-person interaction, helps in individuating the action of a person by suggesting a context; the second, person-person interaction, identifies a group activity.

These approaches suffer from lack of generalization: they focus on a restricted set of actions, which are specific for a particular scenario. In this sense, a versatile generative model is presented in [33], where interacting events in crowded scene are modelled in an unsupervised way, and interactions are modeled as co-occurrences of atomic events. No tracking is performed due to the high people density, and local motions are considered as low-level features instead.

Approaches where sociological aspects are taken into account are [3, 19, 26, 27, 29, 31]. The keystone model that explains and simulates the human dynamics in crowd as a gas-kinetic phenomenon is the social force model (SFM) [10]. Here, interacting means being close each other during a walk or a run, and is explained as a balance between repulsive and attractive terms. The social force model has been modified in [27], where SFM is embedded as model for the dynamics in a tracking framework. Independently, a variational learning strategy is proposed in [31], where a dynamic model is trained for predicting the position

of moving subjects, employing the SFM. In [26], a versatile synergistic framework for the analysis of multi-person interactions and activities in heterogeneous situations is presented. An adaptive context switching mechanism is designed to mediate between two stages, one where the body of an individual can be segmented into parts, and the other facing the case where people are assumed as rigid bodies. The concept of spatio-temporal personal space is also introduced to explain the grouping behavior of people. They extend the notion of *personal space* [2] to that of *spatio-temporal personal space*. Personal space is the region surrounding each person, that is considered personal domain or territory. Spatio-temporal personal space takes into account the motion of each person, modifying the geometry of the personal space into a sort of cone. This multi-person interaction approach share some similarities with our proposal, however, the sequences presented in the paper show very few people (max 3), and simpler situations. A quite novel perspective for detecting interactions in video surveillance scenarios come from the estimation of the human gaze (i.e., the head direction) in low resolution images [28]: in [3] the head direction serves to infer a 3D visual frustum as approximation of the focus of attention (FOA) of a person. Given the FOA and proximity information, interactions are estimated: the idea is that close-by people whose view frustum is intersecting are in some way interacting. In the experiments, we compared with this approach, abbreviated as IRPM. The same idea has been explored, independently, in [29]. Our approach improves this intuition, studying more in detail how people are usually located w.r.t each other during the interaction.

3 Our approach

An F-formation can be specified by the related o-space and the oriented positions of the participants. Suppose we know the oriented positions of the subjects in the scene on the ground plane. Our algorithm jointly estimates the o-space(s) and the subjects involved in the related F-formation(s). The main idea is sketched through the toy example of Fig. 2a-c. Let us focus on $K = 2$ subjects, i and j, located at positions (x_i, y_i) and (x_j, y_j) with head orientation α_i and α_j, respectively. They are exactly facing each other, as depicted by the dashed line connecting their heads (Fig. 2a). Let us also suppose they are at a distance where social interaction can take place, i.e., $d = 1.5$ meters [1]. Given these (hard) constraints, each k-th subject votes for a candidate center $C(k)$ of the o-space, which has coordinates $x_{C(k)}, y_{C(k)}$:

$$C(k) = \left[x_{C(k)}, y_{C(k)}\right] = \left[x_k + r \cdot cos(\alpha_k), y_k + r \cdot sin(\alpha_k)\right], \quad k = 1, \ldots, K \tag{1}$$

where the radius $r = d/2 = 0.75$. Each vote is accumulated in an *intensity accumulation space* $\mathcal{A}_I$, at entry $\tilde{x}_{C(k)}, \tilde{y}_{C(k)}$, where the tilde refers to the closest integer approximation (opportunely rounding the real value resulting from Eq. (1)) determined by the discretisation of the space $\mathcal{A}_I$. At the same time, the *ID labels* i and j are stored at the same entry of a *label accumulation space* $\mathcal{A}_L$, having the same size of $\mathcal{A}_I$. In the toy example of Fig. 2a, both people vote for a coincident location (Fig. 2b), which becomes the center of a *candidate* o-space (Fig. 2c).

To recover the subjects related to this candidate o-space, it is sufficient to access the labels in $\mathcal{A}_L$ associated to the votes in that location. We now know that the center of the candidate o-space has been voted by subjects i and j. At this point, the important condition

[1] We will discuss this assumption later in the experiments.

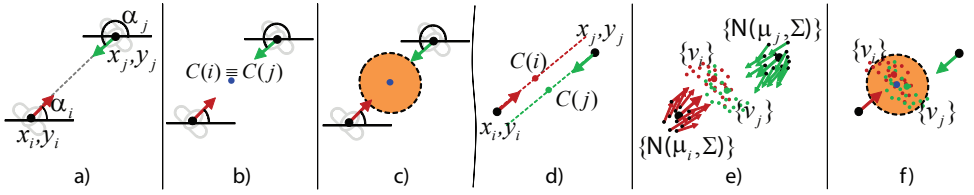


Figure 2: Scheme exemplifying the proposed approach. (a-c) Two subjects exactly facing each other at a fixed distance vote for the same center of the circumference representing the o-space. (d) The 2 subjects do not face each other exactly in real cases. (e-f) Several positions and head orientations are drawn from Gaussian distributions associated to the subjects so as to deal with the uncertainty of real scenarios, robustifying the proposed approach.

of *"no-intrusion"* should be checked for the sociological consistence of the candidate o-space. The no-intrusion condition states: a candidate o-space for the subjects i and j does not have to contain other subjects different from i and j. If the no-intrusion condition is fulfilled the candidate o-space becomes a *valid* o-space.

One could object that the scenario depicted in Fig. 2a-c would be very rare. In fact, our experiments on real data suggest that people engaged in a discussion are rarely positioned on an exact circumference and facing its center. Moreover, computer vision methods are still not capable of estimating head orientation with high precision, and only a coarse quantization of this angle is typically considered in the current state of the art [4]. These two facts make the above deterministic, hard scheme ineffective. For example, no candidate o-space would be detected for the case in Fig. 2d where the subjects do not lie on the same diameter.

In order to deal with this problem, we inject uncertainty in the voting procedure, proposing an algorithm which is sketched in Fig. 2e-f. The proposed procedure is structured in three distinct stages and in the following we present an explanation for each step[2].

Sampling. We assume the positions and the (head) orientation of the different subjects as uncertain to some extent and modeled as random Gaussian variables, i.e.,

$$[x_k, y_k, \alpha_k]^T \sim \mathcal{N}(\mu_k, \Sigma_k) \tag{2}$$

where $\mu_k = [x_k, y_k, \alpha_k]^T$ and $\Sigma_k = \Sigma = \text{diag}(\sigma_x^2, \sigma_y^2, \sigma_\alpha^2)$. We transfer this uncertainty in the voting approach by drawing $N-1$ (being μ_k the N-th sample) i.i.d samples from every k-th distribution[3], as depicted in Fig. 2e. Each n-th sample of the k-th subject $s_{n,k} = [x_{n,k}, y_{n,k}, \alpha_{n,k}]^T$ has associated a weight $w_{n,k}$, which is the likelihood of being extracted from its generating distribution, i.e., $w_{n,k} = \mathcal{N}(s_{n,k}|\mu_k, \Sigma)$ and a label $l_{n,k} = k$, that links it to the related k-th individual.

Voting. Each sample votes for a candidate position in the same way of Eq. 1. The vote in the accumulation space $\mathcal{A}_I$ given by the n-th sample with weight $w_{n,k}$ adds $w_{n,k}$ in the accumulator, thus modeling the uncertainty associated to that sample. In this way, the accumulation space grows in number of votes, which are sparsely distributed. The accumulation of identity labels in $\mathcal{A}_L$ is done similarly for each sample as explained for the toy example in Fig. 2.

[2]Additional material at http://profs.sci.univr.it/~cristanm/publications.html includes a pdf with a summary of the algorithm as a scheme.

[3]In this paper, we fix Σ and the number of samples for all the people observed. However, interesting policies can be adopted in dependence on the certainty we have in the k-th subject (for example due to the tracker providing the subject position, or to the classifier estimating the head orientation).

Once the accumulation process is finished, the matrix $\mathcal{A}_I$ is revised with $\tilde{\mathcal{A}}_I$:

$$\tilde{\mathcal{A}}_I(x,y) = \text{card}(x,y) \cdot \mathcal{A}_I(x,y) \quad \text{for each } x,y \in \mathcal{A}_I(x,y) \tag{3}$$

where $\text{card}(x,y)$ counts the different subjects that voted in $\mathcal{A}_I(x,y)$. Such information is easily extracted from $\mathcal{A}_L(x,y)$. In this way, a high vote is given in those positions that have been voted with strong weights by many subjects. After that, the o-space may be found by looking for the maximum values of $\tilde{\mathcal{A}}_I$, and the associated subjects can be identified by checking $\mathcal{A}_L$.

O-space validation. The evaluation of the no-intrusion condition is performed by analyzing how strong is the presence in the o-space of an external subject. Following a probabilistic approach, we compute the maximum weight $w^*_{n,h}$ of a sample of an external subject h which falls in the candidate o-space. A high $w^*_{n,h}$ in an o-space of center (x_c, y_c) mirrors a high probability that h is invading that o-space. A threshold τ_{INTR} is used to detect the invading external subject. If this happens, the o-space is invalid, and the intensity accumulator is updated imposing $\tilde{\mathcal{A}}_I(x_c, y_c) = 0$, and the search for the maximum value on the updated $\mathcal{A}_I$ is repeated.

This algorithm extends naturally to F-formations composed by more than two subjects and to more F-formations in the same scene thanks to the characteristics of the Hough voting scheme. Actually, in a crowded situation, there could easily be more than one F-formation. Thus, we need to check all the possible o-spaces efficiently, and this is done in the following way. Consider the case of two subjects i and j with their o-space detected as described in the *O-space validation* stage. The accumulators $\mathcal{A}_I$ and $\mathcal{A}_L$ are then updated by pruning away the votes given by $\{w_{n,i}\}$ and $\{w_{n,j}\}$ in $\mathcal{A}_I$, respectively, and removing the labels i and j from $\mathcal{A}_L$. Then, $\tilde{\mathcal{A}}_I$ is re-computed. The max search process on $\tilde{\mathcal{A}}_I$ and the no-intrusion check are thus repeated, and this is iterated until no more o-spaces are found. This strategy has also the beneficial effect of providing the F-formations in decreasing order of likelihood, assuming the likelihood of an F-formation proportional to the accumulation of votes (which can be assimilated to probabilities) in the center of the related o-space stored in $\mathcal{A}_I$.

4 Experiments

Our algorithm has been tested on synthetic and real data. The former proves the effectiveness of our algorithm in detecting groups disregarding a-priori errors due to bad tracking or wrong head orientation estimations. The latter considers two different real scenarios, one indoor and one outdoor, where errors may occur.

As accuracy measures, we estimate that a group has been correctly estimated if at least $\lceil (2/3 \cdot |G|) \rceil$ of their components are found, where $|G|$ is the cardinality of group G. This rule has an exception that holds in the case $|G| = 2$. In that case, all the components must be detected. Given this, for each situation analyzed we estimate the *precision* and *recall* of finding groups, averaged over time.

In addition, to further promote the versatility of our framework, we build for each sequence a *relation matrix* P_2 that represents how many times two people stand in the same group for a certain period of time. Actually, during a party, people may change groups, standing alone for a while, re-joining a conversation, etc.. P_2 analyzes the strength of pairwise relations and, for example, is capable to indicate, given a person, who is the subject with which she/he is interacting most. This matrix has been employed in other social signalling techniques [6], and we can compare it with the analogous matrix built employing

the ground-truth data. A measure of the similarity between the two matrices has been performed employing the *Mantel Test* [23], which is commonly used in cluster analysis to test the correlation between two distance matrices. It operates by evaluating correlations scores from repeated randomizations of the entries of the matrices. If randomizations frequently produce a correlation stronger or as strong as the original data, there is little evidence that the correlation between the two matrices differs from zero. In rough terms, it is a measure of similarity between matrices which actively takes into account their structure.

The proposed method is compared with the Inter-Relation Pattern Matrix method[4] (IRPM) proposed in [3], whose description is reported in Sec. 2.

The free parameters of the method are the radius r, the variances $\sigma_x^2, \sigma_y^2, \sigma_\alpha^2$, the number of samples per-person N, and the threshold τ_{INTR} of the no-intrusion condition. Choosing such values is very intuitive, and it can be driven by sociological and empirical considerations. As an example, the setting of the radius r is a matter of pure sociological aspects: Hall [9] defines 4 relational ranges of distances that witness the type of relation a subject has with the others, and are (expressed in meters): $[0, 0.45]$ for *intimate* relations, $(0.45, 1.2]$ for *casual/personal* relations, $(1.2, 3.5]$ for *social/consultive* relations, and > 3.5 for no-relation. Now, suppose that two people are involved in a vis-a-vis interaction. They may make a circular o-space whose diameter is $2r$. In all the other F-formations, the distance among two people is $< 2r$. Therefore, r represents half of the maximal distance two people may lie in the space and being judged as connected in an F-formation. If we set $r = 60cm$, we are interested in a upper bound that becomes the *casual/personal* range, because $2r = 120cm$

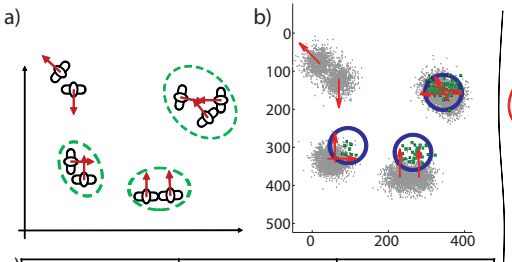
The parameters σ_x^2 and σ_y^2 allow to project the position of the people in different positions, covering a range of $3\sigma_{x(y)}$. In other words, these values allow to be flexible about the classes of relations taken into account by the r parameter. We fix $\sigma = \sigma_x^2 = \sigma_y^2 = 400cm$, considering thus a range of maximal distances for the F-formations of $2[r - 3\sigma, r + 3\sigma] = [0, 240]cm$. The value of σ_α^2 depends on the quantization of the head orientation. We employ 4 head orientations, so $\sigma_\alpha^2 = 0.005$ is a reasonable value. The parameter N can be instead chosen by considering computational aspects. In the current, non-optimized MATLAB version it takes averagely 15 second per frame using $N = 800$.

Finally, the last parameter τ_{INTR} checks the weights (i.e. likelihood probabilities) of the intruder samples. Therefore, its setting mirrors how tolerant we want to be in considering a sample as a genuine representative of an intruder, depending on its weight. We fix $\tau_{INTR} = 0.7$. Once the parameters are set, they are kept fixed for all the experiments.

4.1 Synthetic data

A psychologist provided 100 different *situations*, where some subjects take part in an F-formation and other do not (examples in Fig. 3d). The input of the tested algorithms is the actual position and head orientation of each subject. The data has been annotated to obtain ground truth of the F-formations. We apply our algorithm and IRPM to all the situations, averaging the precision and the recall scores of all the situations. Fig. 3a shows an examplar situation from the synthetic dataset. Fig. 3b depicts how the sampling process propagates instances of a subject in gray, the votes of the intensity accumulator in green, and the resulting o-spaces in blue. A qualitative analysis has been reported in Fig. 3b. The ground truth is depicted in dotted green, whereas the results of our approach and IRPM are in blue and in red, respectively. Our approach is able to model interactions where IRPM fails. In case (iii), our approach fails in estimating the two vis-a-vis interaction, being them very close. In

[4]The code is available at http://www.lorisbazzani.info/code-datasets/irpm/



c)

Method	precision	recall
IRPM	0.7130	0.5389
Our approach	**0.75**	**0.8565**

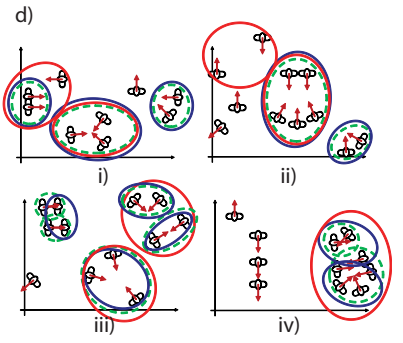


Figure 3: Experiments with synthetical data (see text). The figure is better viewed in color.

c)

d)

Seq.	IRPM	our approach
1	p=0.63|r=0.54|M=0.62	**p=0.66|r=0.67|M=0.81**
2	p=0.55|r=0.19|M=0.67	**p=0.85|r=0.57|M=0.76**

e)

1,2,3,4,5,6,10	p=0.80|r=0.55|M=0.49	**p=0.89|r=0.57|M=0.87**

Figure 4: Experiments with real data (see text). In the tables, p, r, M stand for (mean) precision, recall, and Mantel score, respectively.

general, looking at the global results in Fig. 3c, one can note that our proposal gets higher rates for both precision and especially for the recall.

4.2 Real scenarios

The outdoor situation is represented by a novel dataset, dubbed *CoffeBreak* and downloadable at http://profs.sci.univr.it/~cristanm/datasets.html. It represents a coffee-break scenario of a social event that lasted 4 days, captured by two cameras. The dataset is part of a social signaling project whose aim is to monitor how social relations evolve over time. Nowadays, only 2 sequences of a single day of a single camera have been annotated, each one covering a period of averagely 1 minute. A psychologist annotated the videos indicating the groups present in the scenes, for a total of 45 frames for *Seq1* (a frame in Fig. 4a-b) and 75 frames for *Seq2* (see Fig. 4c). The annotations have been done by analyzing each frame and a set of questionnaires that the subjects filled in. The dataset is still challenging from the tracking and head pose estimation point of view, due to multiple occlusions. This enables us to test our technique in a very noisy situation.

Since CoffeBreak is a crowded scenario, occlusions make extremely hard full human bodies detection. Thus, the subjects' heads are the only cues to perform tracking in a robust way. To extract the head locations of all the subjects in the scene we adopted a system based on class-specific Hough forests [7] trained on human heads in all possible orientations. This allowed us to reliably detect all the possible head candidates in the scene, independently from their orientation with respect to the ground plane. After performing head detection in all the frames, such detections needed to be filtered and linked in order to generate plausible ground plane trajectories of all the subjects. To this end, the ground plane homography and an estimation of the average height of the subjects were used to compute the ground plane location corresponding to each head detection. Consecutive detections corresponding to the same subject were linked by matching appearance descriptors. Finally, head orientation

detection has been performed on 4 classes employing the covariance based approach of [32] (see Fig. 4a). Once the oriented positions of the head are given, we estimate the ground plane homography given a set of measurements obtained on site.

The mean precision, recall score and the Mantel correlation reported in Fig. 3d show that our approach outperforms IRPM. In Fig. 3a-b some qualitative results are depicted: in Fig. 3a we have the head detection results together with the orientation. In Fig. 3b, the blue segments indicate the groups found by our approach (the ground truth is (6,7),(11,12,5),(3,10)). IRPM did not find any groups in that frame.

The indoor data come from a publicly available dataset for group detection, called GDet 2010 and downloadable at http://www.lorisbazzani.info/code-datasets/multi-camera-dataset/. The dataset is made by 12 subsequences of about 2 minutes each, with the availability of the full camera calibration parameters. GDet 2010 videos consider a vending machines area where people take coffee and other drinks, and chat in the spare time. The videos have been acquired with two monocular cameras, located on opposite angles of a room close to the floor. People involved in the experiments were not aware of the aim of the trials and behaved naturally. The ground truth has been made by a psychologist like in the CoffeeBreak scenario. Afterwards, some of them were asked to fill in a form inquiring if they talked to someone in the room and to whom. The videos have been analyzed by a psychologist, that noted the social exchanges occurred and produced the ground truth of social interactions. In this case, people tracking has been performed using Hybrid Joint-Separable (HJS) filter proposed in [20], for its capability of dealing with occlusions by means of the estimation of the occlusion maps exploiting the camera calibration. Given the bounding boxes of the tracked people, the head is approximately located within a bounding box. Then, head pose estimation is performed like in the CoffeeBreak scenario.

A quantitative analysis of the results on a subset of sequences is reported in Fig. 4e. Even in this case, our approach outperforms IRPM. Note the values of the Mantel tests: in general, our approach draws a social situation in terms of pairwise relations which is close to the ground truth.

5 Conclusions

This paper presents a sociologically principled method for the detection and analysis of human interactions exploiting F-formations. An F-Formation is a plausible ensemble of possible spatial and orientational organisation people assume during the course of an interaction. Our approach aims at automatically detecting the main social space identified by the sociological findings, the so called o-space, which is a space internal to the interacting people in which no other people are allowed to lie. The net result is a brand new robust interaction detection algorithm based on a well-established sociological theory able to deal with simple to moderately crowded scenes.

The approach has been tested on synthetic data and real scenarios proving its robustness and accuracy in the disparate situations addressed. This is appreciable per se (as compared to ground truth) and also ameliorates the current state of the art results of the IRPM-based method. These results are obtained dealing with complex scenario in which the people detection, the orientation of their heads, and tracking are difficult, likely producing inaccurate input data. Still, our algorithm performs quite well in detecting interactive groups thanks to the statistical voting process.

So far in the literature, this is the first approach that discovers social interactions based on

the automatic detection of F-formations solely from visual cues. Many improvements can be certainly envisaged for the future work. From the algorithmic point of view, clear and obvious improvements may derive from the use of the temporal information provided by the tracking, so as from the adoption of more reliable and efficient people detection and head orientation classification methods. From the application perspective, additional features extracted from the detected F-formations may support the comprehension of the interactions, possibly predicting the likely outcome, which can be useful in evaluating situations of social interest.

Acknowledgments

Research funded by the EU-Project FP7 SAMURAI, grant FP7-SEC-2007-01 No. 217899.

References

[1] J.K. Aggarwal and S. Park. Human motion: Modeling and recognition of actions and interactions. In *Proc. 2nd International Symposium 3D Data Processing, Visualization, and Transmission*, pages 640–647. IEEE Computer Society, 2004.

[2] I. Altman. *The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding.* Brooks/Cole Publishing Company, Monterey (CA), 1975.

[3] L. Bazzani, M. Cristani, D. Tosato, M. Farenzena, G. Paggetti, G. Menegaz, and V. Murino. Prai*hba special issue: Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems, The Journal of Knowledge Engineering*, 2011. in print.

[4] L.M. Brown and Y.L. Tian. Comparative study of coarse head pose estimation. In *Proc. Motion and Video Computing Workshop*, pages 125–130, 2002.

[5] Z. Cheng, L. Qin, Q. Huang, S. Jiang, and Q. Tian. Group activity recognition by gaussian processes estimation. In *Proc. International Conference on Pattern Recognition (ICPR)*, pages 3228 –3231, 2010.

[6] L. Freeman. Social networks and the structure experiment. In *Research Methods in Social Network Analysis*, pages 11–40, 1989.

[7] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009.

[8] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: a review. *Image and Vision Computing*, 27:1775–1787, 2009.

[9] E.T. Hall. *The hidden dimension*. Doubleday New York, 1966.

[10] D. Helbing and P. Molnár. Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282–4287, 1995.

[11] S. Hongeng and R. Nevatia. Large-scale event detection using semi-hidden markov models. In *Proc. IEEE International Conference on Computer Vision*, volume 2, 2003.

[12] Y.A. Ivanov and A.F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:852–872, August 2000.

[13] T. Jebara and A. Pentland. Action reaction learning: Automatic visual analysis and synthesis of interactive behaviour. In *Proc. First International Conference on Computer Vision Systems*, ICVS '99, pages 273–292. Springer-Verlag, 1999.

[14] A. Kendon. *Studies in the Behavior of Social Interaction*. Lisse: Peter De Ridder Press, 1977.

[15] A. Kendon. Some theoretical and methodological aspects of the use of film in the study of social interaction. *Emerging strategies in social psychological research*, pages 67–91, 1979.

[16] A. Kendon. *Conducting Interaction: Patterns of behavior in focused encounters*. Cambridge University Press, 1990. ISBN 0521389380.

[17] A. Kendon. Spacing and orientation in co-present interaction. In *Development of Multimodal Interfaces: Active Listening and Synchrony*, volume 5967 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin / Heidelberg, 2010.

[18] H. Kuzuoka, Y. Suzuki, J. Yamashita, and K. Yamazaki. Reconfiguring spatial formation arrangement by robot body orientation. In *Proc. 5th ACM/IEEE International Conference on Human-robot interaction*, pages 285–292, New York, NY, USA, 2010. ACM.

[19] T. Lan, Y. Wang, W. Yang, and G. Mori. Beyond actions: Discriminative models for contextual group activities. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[20] O. Lanz. Approximate bayesian multibody tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(9):1436 –1449, 2006.

[21] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *In ECCV workshop on statistical learning in computer vision*, pages 17–32, 2004.

[22] W. Lin, M.T. Sun, R. Poovendran, and Z. Zhang. Group event detection with a varying number of group members for video surveillance. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(8):1057 –1067, 2010.

[23] N. Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2):209, 1967.

[24] B. Ni, S. Yan, and A.A. Kassim. Recognizing human group activities with localized causalities. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1470–1477, 2009.

[25] N. Oliver, B. Rosario, and A. Pentland. Graphical models for recognising human interactions. In *Advances in Neural Information Processing Systems*, 1998.

[26] S. Park and M.M. Trivedi. Multi-person interaction and activity analysis: a synergistic track- and body-level analysis framework. *Mach. Vision Appl.*, 18:151–166, 2007.

[27] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: modeling social behavior for multi-target tracking. In *Proc. 12th International Conference on Computer Vision, Kyoto, Japan*, pages 261 –268, 2009.

[28] N. Robertson and I. Reid. Estimating gaze direction from low-resolution faces in video. In *Proc. European Conference on Computer Vision*, volume 3952, pages 402–415. 2006.

[29] N.M. Robertson and I.D. Reid. Automatic reasoning about causal events in surveillance video. *EURASIP Journal on Image and Video Processing*, 2011.

[30] R.J. Rummel. *Understanding conflict and war*. Sage Publications, 1981.

[31] P. Scovanner and M.F. Tappen. Learning pedestrian dynamics from the real world. In *Proc. International Conference on Computer Vision*, pages 381–388, 2009.

[32] D. Tosato, M. Farenzena, M. Spera, V. Murino, and M. Cristani. Multi-class classification on riemannian manifolds for video surveillance. In *Proc. European Conference on Computer Vision*, volume 2, pages 378–391, 2010.

[33] X. Wang, X. Ma, and W.E.L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31:539–555, March 2009.