Socially-Driven Computer Vision for Group Behavior Analysis

Marco Cristani and Vittorio Murino

Abstract. The analysis of human activities is one of the most intriguing and important open issues in the video analytics field. Since few years ago, it has been handled following primarily Computer Vision and Pattern Recognition methodologies, where an activity corresponded usually to a temporal sequence of explicit actions (run, stop, sit, walk, etc.). More recently, video analytics has been faced considering a new perspective, that brings in notions and principles from the social, affective, and psychological literature, and that is called Social Signal Processing (SSP). SSP employs primarily nonverbal cues, most of them are outside of conscious awareness, like face expressions and gazing, body posture and gestures, vocal characteristics, relative distances in the space and the like. This paper will discuss recent advancements in video analytics, most of them related to the modelling of group activities. By adopting SSP principles, an age-old problem -what is a group of people?- is effectively faced, and the characterization of human activities in different respects is improved.

Introduction

Detecting human interactions represents one of the most intriguing frontiers of the automated surveillance since more than a decade. Recently, sociologic and psychological findings have been considered into video surveillance algorithms, especially thanks to the advent of Social Signal Processing work, a recent multi-disciplinary area where computer vision and social sciences converge. This chapter follows this

Marco Cristani
Università degli Studi di Verona and Pattern Analysis & Computer Vision, Istituto Italiano di Tecnologia, via Morego, 30, 16163 Genova
e-mail: marco.cristani@iit.it

Vittorio Murino
Pattern Analysis & Computer Vision, Istituto Italiano di Tecnologia, via Morego, 30, 16163 Genova
e-mail: vittorio.murino@iit.it

R. Cipolla et al. (eds.), *Registration and Recognition in Images and Videos*,
Studies in Computational Intelligence 532,
DOI: 10.1007/978-3-642-44907-9_10,

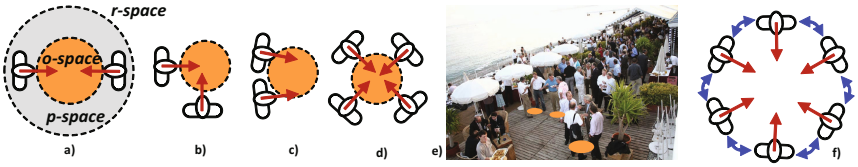


Fig. 1 F-formations: a-d) The component spaces of an F-formation: vis-a-vis, L, side-by-side, and circular F-formations, respectively. O-spaces are drawn in orange. e) Cocktail-party scene where some o-spaces are superimposed in orange.

direction and proposes a detailed overview on our recent activity on the analysis of group activities. In particular, we will present three scenarios where a group of interacting people is first detected, using positional and orientation features [15]; subsequently, the group is characterized by inferring the social relations between the participants exploiting proxemics cues [18]; finally, voice activity is detected by employing solely visual cues [16].

1 Analysis of Social Interactions Using F-formations

The first contribution is devoted to detect social interactions using statistical analysis of spatial-orientation arrangements that have a sociological relevance ([15]). As social interactions we intend the acts, actions, or practices of two or more people mutually oriented towards each other; more in general, any dynamic sequence of social actions between individuals (or groups) that modify their actions and reactions by their interaction partner(s). 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 notion of F-formation as defined by Adam Kendon in the late 70s ([39]).

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 [44].

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 [47], 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 Section is organized as follows. In Sec.1.1, a review of the literature concerning the interaction modelling in surveillance settings is given. The proposed approach is detailed in Sec. 1.2, and the experiments are reported in Sec. 1.3. Finally, Sec. 1.6 concludes the paper with remarks and a discussion on the several possible future developments.

1.1 Group Interaction Discovery: State of the Art

A dated but interesting review on methods that consider human interactions is presented in [2], that focuses especially on motion cues. Pioneering studies on interactions focus on two-agent behaviors, employing statistical learning [56], a mix between syntactical and statistical pattern recognition paradigms [36], or Action-Reaction Learning [37]. 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 [24] for a comprehensive review. Moving to unconstrained scenarios, as those typical of the videosurveillance field, the spectra of the activities modeled becomes narrower. In [34] 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 [55, 12], 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 [48], 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 [45], 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 [72], 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 [59, 65, 58, 45, 62, 6]. The keystone model that explains and simulates the human dynamics in crowd as a gas-kinetic phenomenon is the social force model (SFM) [32]. 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 [59], where SFM is embedded as model for the dynamics in a tracking framework. Independently, a variational learning strategy is proposed in [65], where a dynamic model is trained for predicting the position of moving subjects, employing the SFM. In [58], 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* [3] 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 [61]: in [6] 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 [62]. Our approach improves this intuition, studying more in detail how people are usually located w.r.t each other during the interaction.

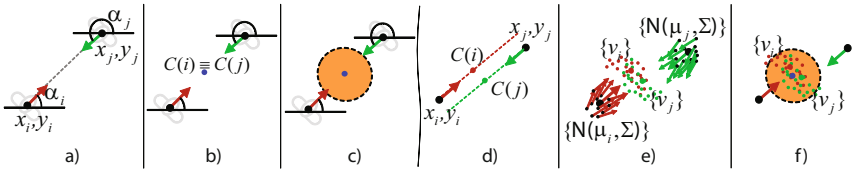


Fig. 2 Scheme exemplifying the proposed approach. (a-c) Two subjects exactly facing each other at a fixed distance vote for the same center of the circumference representing the o-space. (d) The 2 subjects do not face each other exactly in real cases. (e-f) Several positions and head orientations are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

1.2 A Socially-driven Method for Detecting Group Interactions

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 $x_{C(k)}, y_{C(k)}$:

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

where the radius $r = d/2 = 0.75$. Each vote is accumulated in an *intensity accumulation space* $\mathscr{A}_I$, 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 $\mathscr{A}_I$. At the same time, the *ID labels* i and j are stored at the same entry of a *label accumulation space* $\mathscr{A}_L$, having the same size of $\mathscr{A}_I$. 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 $\mathscr{A}_L$ 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 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.

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

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 [10]. 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 $\Sigma_k = \Sigma = \mathrm{diag}(\sigma_x^2, \sigma_y^2, \sigma_\alpha^2)$. 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 $\mathscr{A}_I$ 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 $\mathscr{A}_L$ is done similarly for each sample as explained for the toy example in Fig. 2. Once the accumulation process is finished, the matrix $\mathscr{A}_I$ is revised with $\tilde{\mathscr{A}_I}$:

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

where $\mathrm{card}(x,y)$ counts the different subjects that voted in $\mathscr{A}_I(x,y)$. Such information is easily extracted from $\mathscr{A}_L(x,y)$. 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 $\tilde{\mathscr{A}_I}$, and the associated subjects can be identified by checking $\mathscr{A}_L$.

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

[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).

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 $\tilde{\mathscr{A}_I}(x_c, y_c) = 0$, and the search for the maximum value on the updated $\mathscr{A}_I$ 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 $\mathscr{A}_I$ and $\mathscr{A}_L$ are then updated by pruning away the votes given by $\{w_{n,i}\}$ and $\{w_{n,j}\}$ in $\mathscr{A}_I$, respectively, and removing the labels i and j from $\mathscr{A}_L$. Then, $\tilde{\mathscr{A}_I}$ is re-computed. The max search process on $\tilde{\mathscr{A}_I}$ 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 $\mathscr{A}_I$.

1.3 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 [22], 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* [50], 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 [6], whose description is reported in Sec. 1.1.

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 [29] 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 $2r = 120cm$

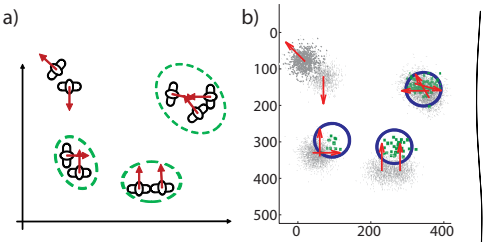
The parameters σ_x^2 and σ_y^2 allow to project the position of the people in different positions, covering a range of $3\sigma_{x(y)}$. In other words, these values allow to be flexible about the classes of relations taken into account by the r parameter. We fix $\sigma = \sigma_x^2 = \sigma_y^2 = 400cm$, 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.

1.4 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

[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**

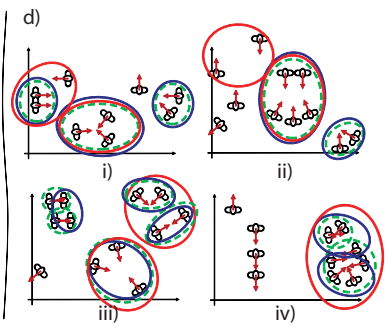
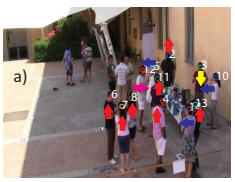
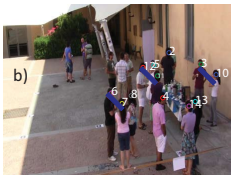


Fig. 3 Experiments with synthetical data (see text). The figure is better viewed in color.





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**

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

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 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.

1.5 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 [23] 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 [69] (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 [46], 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.

1.6 Remarks

This section 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.

2 Inferring Social Relations from Interpersonal Distances

The second contribution is about the characterization of the group interaction aimed at the recognition of the relations among the interlocutors by using proxemic cues [18]. Proxemics can be defined as the "[...] *the study of man's transactions as he perceives and uses intimate, personal, social and public space in various settings* [...]", quoting Hall [30, 31], the anthropologist who first introduced this term in 1966. In other words, proxemics investigates how people use and organize the space they share with others to communicate, typically outside conscious awareness, socially relevant information such as personality traits (e.g., dominant people tend to use more space than others in shared environments [49]), attitudes (e.g., people that discuss tend to seat in front of the other, whereas people that collaborate tend to seat side-by-side [63]), etc..

This section focuses on one of the most important aspects of proxemics, namely the relationship between physical and social distance. In particular, the section shows that interpersonal distance (measured automatically using computer vision techniques) provides physical evidence of the social distance between two individuals, i.e. of whether they are simply acquainted, friends, or involved in a romantic relationship. The proposed approach consists of two main stages: the first is the automatic measurement of interpersonal distances, the second is the automatic analysis of interpersonal distances in terms of proxemics and social relations (see Section 2.3 for details).

The choice of distance as a social relation cue relies on one of the most basic and fundamental findings of proxemics: People tend to unconsciously organize the space around them in concentric zones corresponding to different degrees of intimacy [30, 31]. The size of the zones changes with a number of factors (culture, gender, physical constraints, etc.), but the resulting effect remains the same: the more two people are intimate, the closer they get. Furthermore, intimacy appears to correlate with distance more than with other important proxemic cues like, e.g., mutual orientation [26]. Hence, it is reasonable to expect that the distance accounts for the social relation between two people.

One of the main contributions of the paper is that the experiments consider an ecological scenario (standing conversations) where more than two people are involved. This represents a problem because in this case distances are not only determined by the degree of intimacy, but also by the need of ensuring that every person can participate in the interaction. This leads to the emergence of stable spatial arrangements, called F-formations (see Section 2.1 for more details) [39], that impose a constraint on interpersonal distances and need to be detected automatically. Furthermore, not all distances can be used because, in some cases, they are no longer determined by the degree of intimacy, but rather by geometric constraints. The approach proposed in this work is to consider only the distances between people adjacent in the F-formation (see Section 2.4 for more details) [39].

The other important contribution is that, in contrast with other works in the literature, the radii of the concentric zones corresponding to different degrees of intimacy are not imposed a-priori, but rather learned from the data using an unsupervised approach. This makes the technique robust with respect to the factors affecting proxemic behavior, like culture, gender, etc., as well as environmental boundaries. In particular, the experiments show how the organization into zones changes when decreasing the space at disposition of the subjects and how the unsupervised approach is robust to such an effect.

Standing conversations are an ideal scenario not only because they offer excellent examples of proxemic behavior, but also because they allow one to work at the crossroad between surveillance technologies, often applied to monitor the behavior of people in public spaces, and domains like Social Signal Processing that focus on automatic understanding of social behavior. This is expected to lead, on the long-term, to socially intelligent surveillance and monitoring technologies [17].

The rest of this section is organized as follows. Section 2.1 introduces the main concepts of proxemics, and Section 2.2 provides a brief survey of the state-of-the-

art in computational proxemics. Section 2.3 presents the approach, and Section 2.4 reports the experiments and results. Finally, Section 2.5 draws some conclusions.

2.1 Fundamentals of Proxemics

The wide spectrum of nonverbal behavioral cues displayed during social interactions (facial expressions, vocalizations, gestures, postures, etc.) is well known to convey information about social and affective aspects of human-human interaction (attitudes, personality, emotions, etc.) [60]. Proxemics has shown that the way people use, organize and share space during gatherings and encounters is a nonverbal cue and it conveys, like all other cues, social and affective meaning [42]. This section provides a short description of the main findings of the discipline, with particular attention to phenomena that can be observed in standing conversations, the scenario investigated in the experiments of this work.

From a social point of view, two aspects of proxemic behavior appear to be particularly important, namely interpersonal distances and spatial arrangement of interactants.

The rest of this section focuses on both aspects, including the most important factors that influence them.

2.1.1 Interpersonal Distances

Interpersonal distances have been the subject of the earliest investigations on proxemics and one of the main and seminal findings is that people tend to organize the space around them in terms of four concentric zones associated to different degrees of intimacy:

- *Intimate Zone*: distances for unmistakable involvement with another body (lover or close friend). This zone is typically forbidden to other non-intimate persons, except in those situations where intrusion cannot be avoided (e.g. in elevators).
- *Casual-Personal Zone*: distances established when interacting with familiar people, such as colleagues or friends. This zone is suitable for having personal conversations without feeling hassled. It also reflects mutual sympathy.
- *Socio-Consultive Zone*: distances for formal and impersonal relationships. In this zone, body contact is not possible anymore. It is typical for business conversations, consultation with professionals (lawyers, doctors, officers, etc.) or seller-customer interactions.
- *Public zone*: distances for non-personal interaction with others. It is a zone typical for teachers, speakers in front of a large audience, theater actors or interpersonal interactions in presence of some physical barrier.

In the case of Northern Americans, the four zones above correspond to the following ranges: less than 45 *cm* (intimate), between 45 and 120 *cm* (casual-personal), between 120 and 200 *cm* (socio-consultive), and beyond 200 *cm* (public). While the actual distances characterizing the zones depend on a large number of factors (e.g.,

culture, gender, physical constraints, etc.), the partition of the space into concentric areas seems to be common to all situations.

2.1.2 Spatial Arrangement: The F-Formations

The spatial arrangement during social interactions addresses two main needs: The first is to give all persons involved the possibility of participating, the second is to separate the group of interactants from other individuals (if any). The result are the *F-formations*, stable patterns that people tend to form during social interactions (including in particular standing conversations): "*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*" [39].

In practice, an F-formation is the proper organization of three social spaces (see Figure 1): O-space, P-space and R-space. The O-space (the most important component of an F-formation) is a convex empty space surrounded by the people involved in a social interaction, every participant looks inward into it, and no external people are allowed in this region. The P-space is a narrow stripe that surrounds the O-space and that contains the bodies of the interactants, the R-space is the area beyond the P-space. There can be different F-formations:

- *Vis-à-vis*: An F-formation in which the absolute value of the angle between participants is approximately 180^o, and both participants share an O-space.
- *L-shape*: An F-formation in which the absolute value of the angle between participants is approximately 90^o, and both participants share an O-space.
- *Side-by-side*: An F-formation in which the absolute value of the angle between participants is approximately 0^o, and both participants share an O-space.
- *Circle*: An F-formation where people is organized in a circle, so that the configuration between adjacent participants can be considered as a hybrid between a L-shape and a Side-by-side F-formation.

The same contextual factors that influence the concentric zones described above, affect F-formations as well.

2.1.3 Context Effects on Proxemics

Proxemic behavior is affected by a large number of factors and culture seems to be one of the most important ones, especially when it comes to the size of the four concentric zones described above. In particular, cultures seem to distribute along a continuum ranging from "contact" (when the size of the areas is smaller) to "non-contact" (when the size of the areas is larger) [31]. Further evidence in this sense is proposed in [73], where people from "contact" cultures are shown to approach one another more than the others, and in [66], where the culture effect has been shown to depend on whether one considers shape of territory, size, central tendencies of encroachment, or encroachment variances (the observations were conducted on beaches). In the same vein, interpersonal distances seem to be affected by ethnicity: e.g., black Americans and Mexicans living in the States appear to have different

"contact" tendencies [31, 5]. The effect of culture seems to change when interaction participants have seats at disposition. In this case, people from supposedly "non-contact" cultures tend to seat closer than the others [33]. Furthermore, the seating arrangement seems not to depend on culture [51].

Seating is just one of the many environmental characteristics that can influence the requirements on interpersonal distance and personal space. The literature has investigated the effect of many other characteristics as well, including lighting [1], indoor/outdoor [13], crowding [27] and room size [75, 64, 14]. The work in [1] investigates the effect of lighting with stop-distance techniques: Experimenters get closer and closer to a subject that remains still and says "stop" when she starts feeling uncomfortable. Subjects in bright conditions (600 lx) allow the experimenters to come significantly closer than the subjects in dim conditions (1.5 lx). A similar effect has been observed for the size of the place where people interact: people allow others to come closer in larger rooms [75], when the ceiling is higher [64][14], and in outdoor spaces [13]. The effects of crowding have been studied as well [27]: Social density was increased in a constant size environment for a limited period of time and participants of larger groups reported greater degrees of discomfort and manifested other forms of stress.

2.2 Computational Proxemics: State-of-the-Art

To the best of our knowledge, only a few works have tried to apply proxemics in computing. One probable reason is that current works on analysis of human behavior have focused on scenarios where proxemics do not play a major role or have relied on laboratory settings that impose too many constraints for spontaneous proxemic behavior to emerge (e.g., small groups in smart meeting rooms) [25, 70].

Most of the computing works that can be said to deal with proxemics concern the dynamics of people moving through public spaces. These works typically model repulsive/attractive phenomena by adopting the Social Force Model (SFM) [32]. In particular, the work in [59, 65] improves the perfomance of a tracking approach by taking into account the distance between a subject being tracked and the other subjects appearing in a scene. An attempt to interpret the movement of people in social terms has been presented in [28], where nine subjects (asked to speak among them about specific themes) were left free to move in a $3m \times 3m$ area for 30 minutes. An analysis of mutual distances in terms of the zones described in Section 2.1 allowed to discriminate between people who did interact and people who did not. In a similar way, mutual distances have been used to infer personality traits of people left free to move in a room [76]. The results show that it is possible to predict Extraversion and Neuroticism ratings based on velocity and number of intimate/personal/social contacts (in the sense of Hall) between pairs of individuals looking at one other.

Another frequent application area is social robotics. Early approaches in the domain simply aimed at making robots to respect the personal space of users [54], but more recent works deal with the initiation, maintenance, and termination of social interactions by modulating reciprocal distances, showing that people use similar

proxemic rules when interacting with robots and when interacting with other people [68]. In [9] a generative model has been developed for selecting a set of reactive behaviors that depend on the distance, speed, and sound of interactants. Distance cues are used by the Roboceptionist [53] for recognizing "Present", "Attending", "Engaged", and "Interacting" people at the entrance of the Robotics Institute at Carnegie Mellon University. In [57], a model for human-robot interaction in a hallway is proposed. The idea is to exploit proxemic cues for letting the robot to react properly at the passage of an individual in a narrow corridor. In [43], a user study focuses on the interaction between a human and a robot in a domestic environment. Interactions were analyzed exploiting the four zones and the F-formations introduced in Section 2.1. The researchers found the Personal zone to be the most commonly occupied one and the "vis-à-vis" F-formation to be the most frequent spatial arrangement.

2.3 Detecting a Flexible Set of Socially Meaningful Distances

The proposed approach includes two main stages: the first is the detection of F-formations, and the second is the inference of social relations from interpersonal distances.

2.3.1 Detection of F-Formations

The goal of this stage is to detect F-formations in videos portraying people involved in standing conversations. The first step is to track the people with a fish-eye camera pointing at interactants in a bird-eye view setting (see Figure 5 for an example). This corresponds to a realistic surveillance scenario and allows one to track people with satisfactory precision (tracking has been performed by exploiting a particle filter on each person [4], employing a standard background subtraction algorithm for highlighting the moving objects [67]. The results of our approach that have been obtained with this tracking strategy have been compared with those obtained via manual tracking, showing very similar results). The detection of the F-formations is performed over the output of the tracking step using the approach described in [15]. The output of the F-formations detection algorithm has been validated by hand and it did not produce any error.

F-formations lasting for less than 5 seconds (50 frames in our implementation) have not been taken into consideration in the experiments of this work. The reason is that the next stage of the processing requires the application of a clustering algorithm and 50 frames is a reasonable amount of data needed to avoid the so-called "curse of dimensionality" [20].

2.3.2 Inference of Social Relations

The output of the first stage is a list of pairs where each element includes two subjects that are adjacent in a detected F-formation. Furthermore, the first stage provides the $2D$ position of each subject on the surface of the room. Such data

is accumulated during a time interval (called the "stable period" hereafter) that does not include creation, break or modifications of an F-formation (e.g., no people change their position in the *P*-region). This ensures that during the time interval under analysis all causes that might change the current F-formation are absent. Such causes can be novel people being involved, people leaving, a change in the environmental conditions like rain (people look for a repair), an intruder (e.g., a vehicle passing by and disrupting the F-formation), etc.. The satisfaction of the conditions above is automatically verified by checking that the relative distances between subjects in a F-formation do not change abruptly (i.e., the changes do not exceed a threshold learned automatically from the data).

During the stable period, the approach collects and pools together all pairwise distances between individuals (for a sketch, see Figure 1 (f)). Distances are collected between the centers of mass of the tracked blobs, where each blob corresponds to a separate person. These are shown to distribute according to different modes (see Section 2.4) that should correspond to the concentric zones described in Section 2.1. The modes have been separated via Gaussian clustering by employing the Expectation-Maximization (EM) [19] learning method. The EM employed here is a variation of the original formulation [21]; it is performed by means of a model selection strategy that is injected in the learning stage and that shows several properties that fit well with the situation at hand. First, it allows one to automatically select in an unsupervised way the right number of Gaussian components (in an Information Theory sense). This is a very important aspect, that permits to let the natural separation of the data emerge without human intervention. Second, it deals satisfactorily with the initialization issue, i.e., the Gaussian parameters fit the data realizing a nearly-global optimal fit, minimizing the probability of overfitting (i.e., a Gaussian component that fit only a few data). In addition, the Gaussian clustering takes into account in a principled way the noise due to possibly unprecise tracking, incorporating it as a variance of the measures.

2.4 Testing the Flexible Distances

This section presents experiments and results obtained in this work.

2.4.1 Experimental Setup

The goal of the experiments is to investigate spontaneous standing conversations in a public space, hence the tests have been performed in an outdoor area of size $3m \times 7m$ (see Fig. 5, row (i), column (a)). The area is empty (no physical constraints or obstacles) and two groups of subjects have been invited, in two separate sessions, to move and behave normally through it. The subjects were told that the experiments were aimed at testing a tracking approach and were unaware of the real motivations behind the experiments. During the sessions, the subjects were left alone and no researcher involved in this work was present.

The experiment took place on February 2011, on a sunny day. The area was monitored with a Unibrain Fire-i Digital Camera, on which fisheye optics was mounted.

The camera was located 7 meters above the floor, and it was held to an architectural element of the infrastructure. Therefore, the impact of the capture device onto the ecology of the environment was minimal. The acquisition frame rate was 10 frames per second. After the data acquisition, video data were rectified for correcting the spherical distortion. The two sessions were 15 minutes long for a total of around 20000 frames. One quarter of hour is a duration long enough to collect evidence of pre-existing social relations and short enough to avoid the emergence of new relations. The first session was recorded at 11 AM and the second at 2 PM.

Each session was split into three 5 minutes long segments corresponding to three different experimental conditions:

- Condition 1: the subjects are free to move through the entire area
- Condition 2: the movements of the subjects are restricted to an area of size $3m \times 3.5m$
- Condition 3: the movements of the subjects are restricted to an area of size $1.5m \times 2.0m$

The physical restrictions were represented by lines and marker on the floor. The goal was to measure the effect of the amount of available space on proxemic behavior.

2.4.2 Results of Session 1

The first session involved six subjects (see Fig. 5): two undergraduate students (*a* and *b*), an assistant professor (*c*), and three PhD students working in the same laboratory (*d*, *e*, *f*), two of them working on the same topic (*e* and *d*). The PhD students and the assistant professor were acquainted before the experiment. The undergraduate students are friends, but they never met before the other subjects. In Fig. 5 row (i) we show the results obtained in the longest stable period (subjects free to move in the entire area, see Section 2.3.2), that in this case lasted 108 frames. The image in column (a)-(b) is the last of the period[5]. In that interval, the group was split into three dyads. The histogram in Figure 5-row (i) shows the distribution of the interpersonal distances between members of the same dyad. The application of a clustering approach shows the existence of two modes centered on 48 and 64 *cm*, respectively. The tables in the figure report the fraction of time distances between each pair of adjacent individuals belong to a given mode for each condition, with the value in bold red indicating the highest (most frequent cluster membership) fraction. The two modes seem to account for two of the zones identified by Hall and, not surprisingly, the dyad involving the assistant professor is the only one where the distance belongs with higher probability to the second mode most of the times. This confirms that the higher social distance between the assistant professor and the PhD student results into a physical distance that is higher (on average) than the one between subjects *a* and *b* (who are friends and both undergraduate students), as well as the one between subjects *d* and *e* (who are both PhD students).

[5] The same applies for all the other pictures in the column (a)-(b), i.e., they are the last frames of the corresponding stable period.

In Condition 2 (3×3.5 meters), the longest stable interval (122 frames) corresponds to a circle F-formation, including all subjects (see Fig. 5-row (ii), pictures at left). The clustering of the interpersonal distances of adjacent subjects reveals this time a three-mode distribution with modes at 44, 69 and 99 *cm*, respectively. The first mode accounts for the distance between *a* and *b* (the two undergraduate friends). The second mode accounts for the distances between *c*, *d*, *e* and *f* (the three PhD students and the assistant professor belonging to the same research group). The third mode accounts mainly for the distances between *a* and *e* and between *b* and *c* (the only pairs where the members were unacquainted before the experiments). In this condition too, the physical distances comply with the social information, even though the distance between the assistant professor and the PhD students does not reflect the difference of status.

In Condition 3 (1.5×2 meters), the longest stable period lasted for 914 frames. People form a circular F-formation, giving now rise to four distinct modes in the space of the pairwise distances (see Fig. 5-row (iii)). Once again, two close friends *a* and *b* stand at the closest distances, separated from the rest of the subjects. In particular, subjects *b* and *c* stand at a very high distance if compared to the other measurements. This highlights the separation that holds between subjects that have different status, i.e., the student and the assistant professor.

The variations across the different conditions suggest the following considerations:

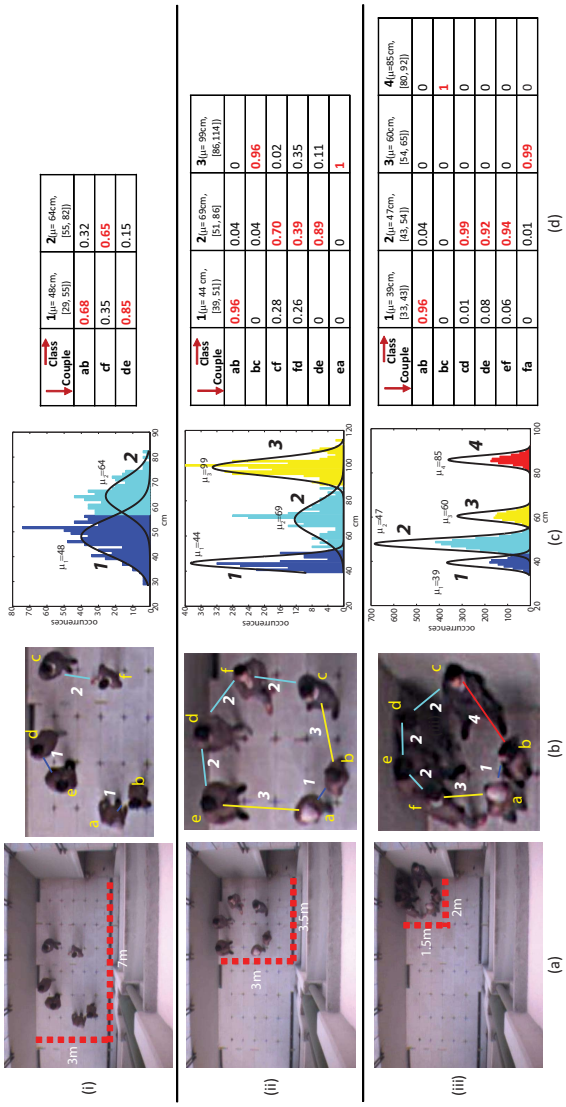
- The histograms show that the modes correspond to shorter distances as the space gets smaller. However, different social relations still result into different modes.
- The fraction of distances that fall in the first mode is 67% in Condition 1, 34% in Condition 2, and 22% in Condition 3.

In other words, the results confirm the findings about the effect of the space at disposition of interpersonal distances and, in particular, the effects of [75] stating that subjects prefer to keep higher distances when the environment gets smaller.

The results shown here analyzed the longest stable period in each session. Anyway, in all the other stable periods, the results were qualitatively similar.

2.4.3 Results of Session 2

The second session involved 7 subjects (see Fig. 6): five undergraduate students acquainted with one another (subjects *a*, *b*, *c*, *d* and *g*), two PhD students that are close friends (subjects *e* and *f*), and the representative of the students in the School of Computer Science (subject *c*).

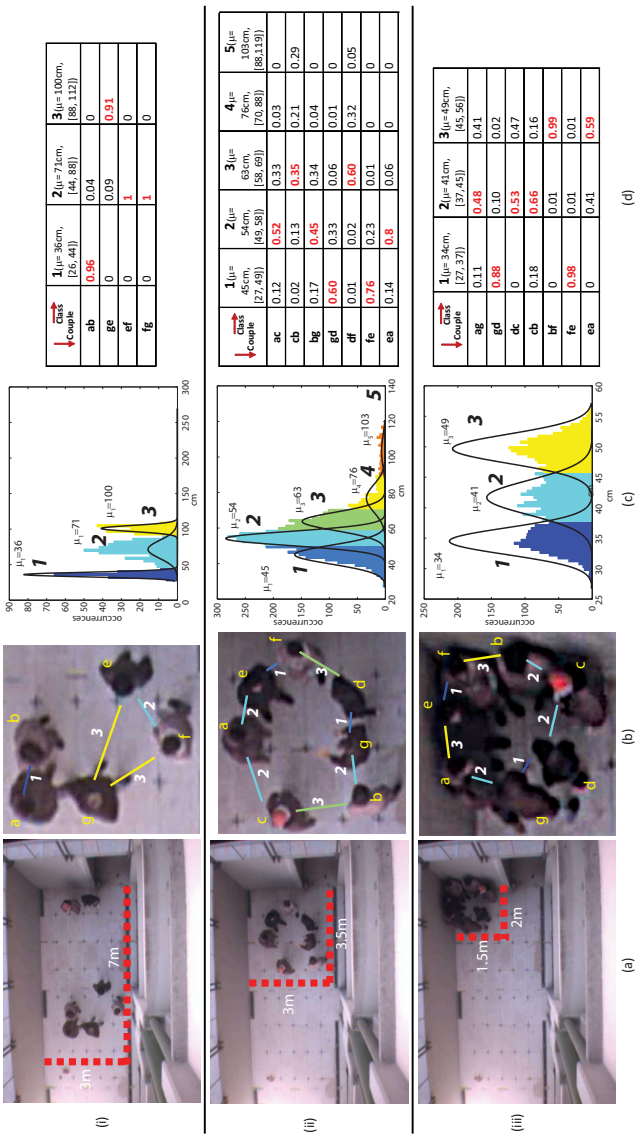


Class / Couple	1(μ= 48cm, [29, 55])	2(μ= 64cm, [55, 82])
ab	0.68	0.32
cf	0.35	0.65
de	0.85	0.15

Class / Couple	1(μ= 44 cm, [39, 51])	2(μ= 69cm, [51, 86]	3(μ= 99cm, [86,114])
ab	0.96	0.04	0
bc	0	0.04	0.96
cf	0.28	0.70	0.02
fd	0.26	0.39	0.35
de	0	0.89	0.11
ea	0	0	1

Class / Couple	1(μ= 39cm, [33, 43])	2(μ= 47cm, [43, 54])	3(μ= 60cm, [54, 65])	4(μ=85cm, [80, 92])
ab	0.96	0.04	0	0
bc	0	0	0	1
cd	0.01	0.99	0	0
de	0.08	0.92	0	0
ef	0.06	0.94	0	0
fa	0	0.01	0.99	0

Fig. 5 The pictures of column (a) show the physical space in which people were free to move. The pictures in column (b) are zoomed versions of those in (a), showing the F-formations detected in each of the three stages. The color of the links corresponds to the color of the most frequent mode to which the distances between the linked individuals belongs to. Rows (i)-(ii)-(iii) refer to Condition 1-2-3, respectively (see text). Histograms in column (c) show the distributions of the distances and the related clustering. The tables in column (d) report the fraction of time distances between each pair of adjacent individuals belong to a given mode. Each mode is identified by the mean, and by the range (in centimeters) of distances it covers (written in squared parentheses). The figure is best viewed in colors.



Class / Couple	1 (μ= 36cm, [26, 44])	2 (μ= 71cm, [44, 88])	3 (μ= 100cm, [88, 112])
ab	0.96	0.04	0
ge	0	0.09	0.91
ef	0	1	0
fg	0	1	0

Class / Couple	1 (μ= 45cm, [27, 49])	2 (μ= 54cm, [49, 58])	3 (μ= 63cm, [58, 69])	4 (μ= 76cm, [70, 88])	5 (μ= 103cm, [88,119])
ac	0.12	0.52	0.33	0.03	0
cb	0.02	0.13	0.35	0.21	0.29
bg	0.17	0.45	0.34	0.04	0
gd	0.60	0.33	0.06	0.01	0
df	0.01	0.02	0.60	0.32	0.05
fe	0.76	0.23	0.01	0	0
ea	0.14	0.8	0.06	0	0

Class / Couple	1 (μ= 34cm, [27, 37])	2 (μ= 41cm, [37, 45])	3 (μ= 49cm, [45, 56])
ag	0.11	0.48	0.41
gd	0.88	0.10	0.02
dc	0	0.53	0.47
cb	0.18	0.66	0.16
bf	0	0.01	0.99
fe	0.98	0.01	0.01
ea	0	0.41	0.59

Fig. 6 The pictures of column (a) show the physical space in which people were free to move. The pictures in column (b) are zoomed versions of those in (a), showing the F-formations detected in each of the three stages. The color of the links corresponds to the color of the most frequent mode to which the distances between the linked individuals belongs to. Rows (i)-(ii)-(iii) refer to Condition 1-2-3, respectively (see text). Histograms in column (c) show the distributions of the distances and the related clustering. The tables in column (d) report the fraction of time distances between each pair of adjacent individuals belong to a given mode. Each mode is identified by the mean, and by the range (in centimeters) of distances it covers (written in squared parentheses). The figure is best viewed in colors.

In Condition 1 (see Fig. 6-row (i)), the group has split into F-formations including $2-3$ people each. Fig. 6 shows the picture of the configuration that has lasted for the longest time (152 frames). The interpersonal distances cluster according to three modes. In the F-formation including three people, the two PhD students (who are close friends) appear to be closer (on average) than the third component (an undergraduate student they are not acquainted with them).

In Condition 2 (see Fig. 6-row (ii)), the most stable configuration is a circle that holds for 629 frames. In this case, the modes are five, but only the first three are used to a significant extent (see the tables of column (d) with the fractions of time distances belong to a given Gaussian component). The two PhD students (e and f) and two undergraduate students (g and d) appear to be closer to one another than the other participants. In the former case, this reflects the fact that they were close friends before the experiment, whereas in the latter, it corresponds to the fact that the two students have a romantic relationship, as it emerged from the questionnaires collected after the experiments. The situation for the other participants is less clear, but this probably happens because all participants are students and their social distances are thus similar. The only factors that seem to make some students closer (see above) are then personal.

In Condition 3 (see Fig. 6-row (iii)), a circular F-formation holds for 592 frames and corresponds to the longest stable interval. There are three modes visibile in the histogram. The PhD students are clearly separated from the rest of the circle (distances belonging to the third mode), while they are very close to one other. The couple (d and g) is tighter than the other dyads as well. In this case again, closer personal relations result into smaller distances.

It is worth to note that the effect of the amount of space at disposition leads to the same conclusions as in session 1 (see end of Section 2.4.2).

2.5 Remarks

This section has presented a study and preliminary experiments on the inference of social relations from interpersonal distances measured automatically via a computer vision approach. The results show that, in accordance with the findings of proxemics, people involved in casual standing interactions tend to get closer when their social relation is more intimate. The experiments have been performed on a limited number of individuals (13 in total), but the setting is fully unconstrained and spontaneous and the results appear to be consistent with the expectations.

An unsupervised analysis of interpersonal distances reveals that the four zones predicted by Hall in his seminal work emerge independently of the space at disposition of the interactants. The radii of the concentric zones are smaller than those measured in [30, 31] for Northern-Americans, but this should not be surprising as the subjects are from Italy, a culture likely to be more "contact" than the American one. Furthermore, the space available to the subjects has been progressively reduced and this has further contributed to reduce the size of the zones. The effects expected

from the reduction of the space have been actually observed, especially when it comes to the tendency to increase interpersonal distances.

The detection of the F-formations appears to be crucial to perform a correct analysis of the interpersonal distances. In fact, previous works in the literature did not consider the geometric constraints imposed by the F-formations and the results have been inconsistent. In contrast, by limiting the analysis only to the distances of neghboring (adjacent) people, our experiments obtain results where social and physical distances match one another.

The next steps to be performed include not only experiments including a larger number of subjects, but also an attempt to use the statistical distributions learned from the data to predict automatically the degree of intimacy between individuals. This would represent a major step towards the development of socially intelligent surveillance technologies.

3 Voice Activity Detection Using Visual Cues in Groups

Following the analysis of non-verbal cues for the detection of social signals, our last contribution is related to the characterization of the group interactions by proposing a Voice Activity Detection approach only based on the automatic measurement of the persons' gesturing activities [16]. This work takes inspiration from the observation that people accompany speech with gestures, the range of visible bodily actions that are, more or less, generally regarded as part of a person‘s willing expression ([40]). Far from being independent phenomena, speech and gestures are so tightly intertwined that every important investigation of language has taken gestures into account, from De Oratore by Cicero (1st Century B.C.) to the latest studies in cognitive sciences ([52, 38]) showing that the two modalities are components of a single overall plan ([40]).

This work presents a method for estimating the level of gesturing as a means to perform Voice Activity Detection (VAD), i.e. to automatically recognize whether a person is speaking or not. The main rationale is that audio, the most natural and reliable channel when it comes to VAD, might be unavailable for technical, legal, privacy related issues or simply for a noisy scenario. A condition that applies in particular to surveillance scenarios where people are monitored in public spaces and are not necessarily aware of being recorded.

Previous works take advantage of restrictive experimental setups in a smart meeting room [35], deploying a system "in the wild" designing a more credible setup for a video surveillance system. We use solely visual cues obtained from only one camera positioned 7 meters above the scene. In particular, the experiments focus on people involved in standing conversations, with an automatic person tracking system that follows each individual. Our VAD method is based on a local optical flow-based descriptor extracted for each individual body, that encodes its energy and complexity using an entropy-like measure. This allows one to discriminate between body oscillations or noise introduced by the tracker, where the optical flow

is low and homogeneous, and genuine gestures, where the movement of head, arms and trunk produces a local flow field which is diverse in both intensity and direction.

The descriptor extracted for each participant produces a signal that can be used for VAD. The proposed approach is interesting for three main aspects. First, the relationship between speech and gestures has been widely documented and studied, but relatively few quantitative investigations of this phenomenon have been made. Second, approaches similar to ours might help to infer information about privacy protected data (speech in this case) from publicly accessible data (gestures in this case): this is also important for establishing whether the simple absence of a certain channel is sufficient to protect the privacy of people and how much. Finally, inferring missing data from available ones can make techniques dealing with challenging scenarios more effective and reliable.

As in the previous section, we suppose to have tracked each individual and additionally to have detected the F-formation. Thus, a square *Region of Interest* (ROI) is defined around each person. The size of the ROI is set automatically to include all gestures of the individual. Areas where multiple ROIs overlap have been ignored to avoid possible confusions between neighboring people.

The measurement technique is applied to each ROI individually and it is expected to accomplish two goals: the first is to discriminate between gestures and postural oscillations typically observed when people stand. The second is to normalize the tracking errors that cause abrupt and spurious shifts of the ROI. The body parts most commonly involved in gesturing are hands, arms, head, and trunk. Their individual movements tend to be very different during gesturing and the measurement values associated to a given ROI try to capture such an aspect:

$$v(t) = \max_{\text{int}}(\{f(t)\}) \times S_{\text{int}}(\{f(t)\}) \times S_{\text{ori}}(\{f(t)\}) \tag{4}$$

where $\{f(t)\}$ is the set of motion flow vectors associated to each pixel of the ROI at time t, $S_{\text{int}}(\{f(t)\})$ is the entropy of the motion flow intensities, and $S_{\text{ori}}(\{f(t)\})$ is the entropy of the orientation values, both calculated over $\{f(t)\}$[6]. The maximum over the flow intensities values $\max_{\text{int}}(\{f(t)\})$ encodes the "energy" associated to the movement, while the two entropic terms serve to highlight those motion flow values which exhibit higher variability in intensity and orientation. In this way, postural oscillations and shifts due to unprecise tracking receive a low score because they cause a global, homogeneous set of intensities and orientations, corresponding to low entropy values. Alternative expressions of $v(t)$ have been considered that use mean and median rather than maximum, or do not include one of the entropy terms. In all cases, the resulting performance is lower than the one obtained with the expression above. A graphical idea of the measurement is given in Figure 7 where colours shift towards red when gesturing activity is higher.

[6] The optical flow has been obtained with the package available at the following URL: `http://server.cs.ucf.edu/~vision/source.html`.

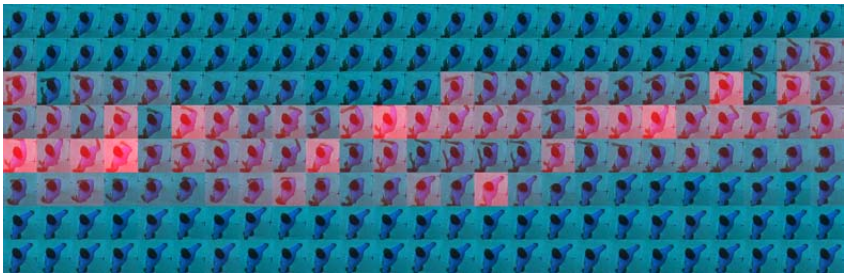


Fig. 7 Qualitative analysis of our descriptor: in the sequence above, an high tonality of red means great gesture activity.

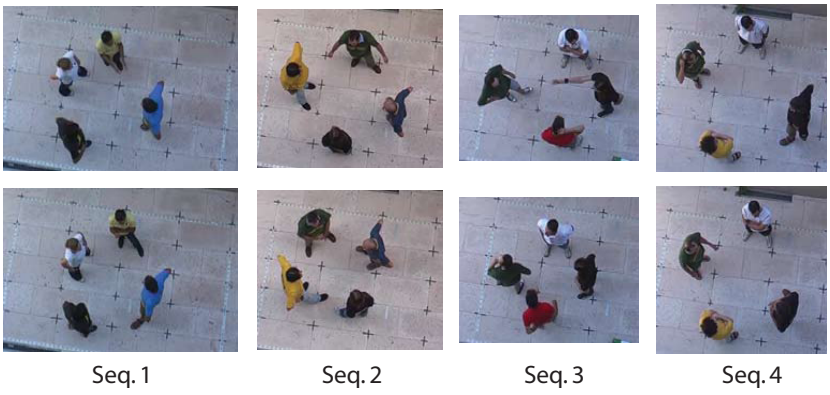


Fig. 8 Some frames of the video sequences used

3.1 Experiments on the Visual VAD

The goal of the experiments is twofold: first, to provide a quantitative measure of the correlation between gestures and speech; second, to measure the effectiveness of the function $v(t)$ (see Section 3.1.2) in a VAD task. Both tasks have been accomplished over *TalkingHeads*, a new dataset publicly available upon request[7] (see some frames in Figure 8).

The dataset contains four conversations lasting, on average, 6 minutes. The data was recorded in a 3.5×2.5 meters wide outdoor area, during a cloudy day in summer. The total number of subjects is 15 (1 female and 14 males), with 4 different participants per conversation (only one subject participated in two conversations). The subjects include 4 academics, 5 undergraduate students, 2 MSc students, 3 post-doctoral researchers, and 1 PhD student. The ages range between 20 and 40 years and the subjects were unaware of the actual goals of the experiments.

[7] `http://profs.sci.univr.it/~cristanm/datasets/TalkingHeads/`

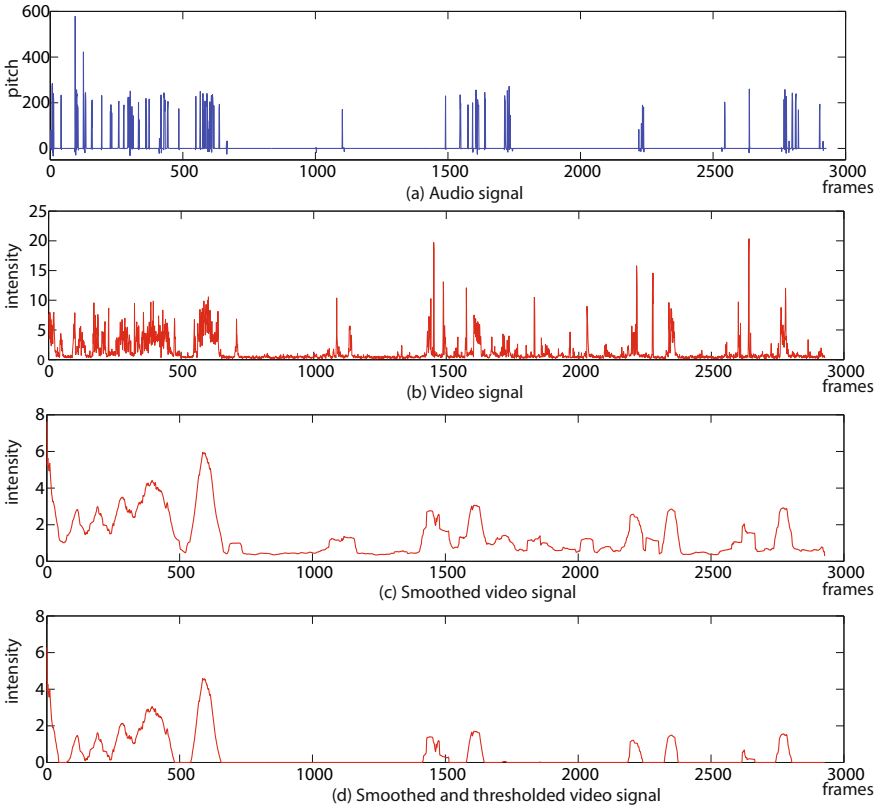


Fig. 9 Examples of signals employed in the analysis. (a) Audio input signal. (b) Video signal produced by our descriptor of a subject involved in the Seq.1. (c) The video signal was smoothed for evaluating the crossmodal correlation (Sec. 3.1.1). (d) The video signal was thresholded for the audio classification (Sec. 3.1.2).

Data were captured at 25 frames per second with a camera positioned 7 meters above the floor and facing downward. The subjects were asked to wear differently colored shirts, in order to make the tracking/localization easier. Tracking has been performed by simple template association. The motion flow has been computed by considering one frame every 4, reducing the video sampling period to 160 *ms*. The audio was recorded at 44100 Hz with 4 wireless headset microphones, each transmitting to its own receiver.

Each audio recording has been segmented into speech and non-speech segments using a robust VAD algorithm based on pitch [41]. This latter was extracted at regular time steps of 10 *ms* with Praat [8], a package including the pitch extraction technique described in [7]. The motivation behind this choice is not only that silence segments are characterized by frequencies way higher than those observed in speech, but also that the pitch tends to be correlated with the "*beat*" gesture

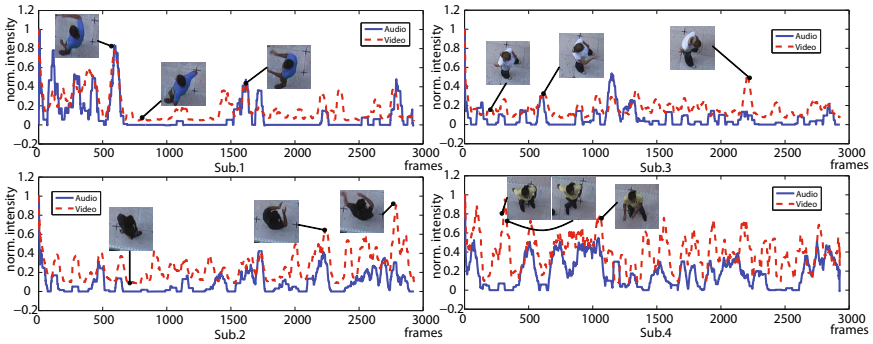


Fig. 10 Visual analysis of the audio and video smoothed data: each plot depicts the smoothed audio (solid blue) and the smoothed video (dashed red) signals for each participant to the dialog. The thumbnails give the feeling of the gesturing activity carried out in a particular instant.

typically accompanying syllables where the intonation is stressed [11, 74]. Then, in order to synchronize audio and video data, audio was resampled according to the video frame rate, averaging the pitch values occurring in each time period. The averaged pitch values constituted the samples of the audio signal that will be analyzed in the following.

3.1.1 Pitch-Gesturing Correlation Analysis

This section shows how the correlation between the pitch (as extracted with Praat), and the gesturing activity (as measured with the approach proposed above) has been measured.

After the application of the techniques described in the previous sections, each sequences results into two signals per person, showing the value of pitch and $v(t)$ at regular time steps of 160 *ms*. Plots (a) and (b) of Figure 9 provide an example of such signals. The simple visual inspection shows that the two signals tend to change according to one another. However, $v(t)$ appears to be more noisy of the pitch because of the sensibility of the optical flow. Hence, both signals have been smoothed with an average filter applied to 8 *s* long windows. Figure 9 (c) shows the smoothed version of $v(t)$, while the smoothed audio and video signals of a complete conversation, normalized with respect to their maximum value, are compared in Fig. 10.

Table 1 reports the Pearson correlation coefficients between $v(t)$ and pitch. Off-diagonal values account for correlations between signals extracted from different individuals. In this way, it is possible to better assess how strong is the correlation between speech and gestures for a given individual.

Table 1 Quantitative measures: correlation coefficients matrix for Seq. 1 . The matrix rows and columns corresponds respectively to the four subsampled video signals (Vsub) and the four subsampled audio signals (Asub) (the non-significant coefficients (p-value$\geq$ 0.05) are underlined in red.

	A sub.1	A sub.2	A sub.3	A sub.4
V sub.1	**0.7310**	0.1338	0.2490	0.0670
V sub.2	0.1900	**0.6454**	0.4460	0.0254
V sub.3	0.1867	0.1966	**0.4838**	-0.0356
V sub.4	-0.2592	0.0472	0.0389	**0.4204**

We performed a similar analysis on the other conversations, with the same parameters, obtaining in total four correlation matrices. Mediating over all the entries in the main diagonal (they were all statistically significant), we obtained a mean correlation score of 0.53, while considering the statistically significant off-diagonals entries we get 0.19. This suggests that $v(t)$ might be a reliable indicator of voice activity. Hence, in the following section, we show how the video signal can be employed to perform VAD.

3.1.2 Voice Activity Detection

The VAD task proposed in this section consists of labeling each frame as *speech* or $non-speech$. As an approximation, each person is treated independently of the others even though the exchange of turns (the opportunity of speaking) tends to follow regularities that might be helpful in improving the performance. The original pitch signal, which has non-zero entries only when the subjects talk, is used as groundtruth.

As a video signal to be used to infer speech, we considered the smoothed signal described above for the correlation analysis. In this way, high frequency components of the original signal have been filtered. The discrimination between speech and non-speech samples has been performed with a thresholding technique. Essentially, as suggested by Fig. 9 and Fig. 10, the video signal has a continuous component caused by small values of optical flow that are always present in the analysis. For this reason, we subtracted the mean to the signal, and we keep the intensities above zero, setting them at 1's. Smoothing and subtraction of the mean represent a thresholding operation that does not need the tuning of any parameter.

At this point, we can compare the two signals, and the detailed analysis of Seq. 1 is shown in Fig. 11.

For the sake of clarity, we report in the figure the (normalized) continuous signals, and not their binary versions which were actually used. As visible, many of the speech samples are correctly captured by the video signal. The figure also reports the precision, recall and accuracy values. In this sequence, the classifier tends to have low recall and high precision (assuming the speech as positive values). Considering

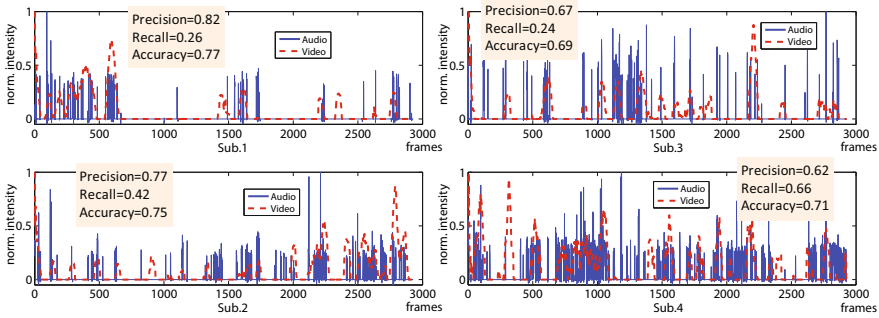


Fig. 11 Audio classification by video analysis. Each plot portrays the audio (solid blue) and the video (dashed red) signals for each participant to the dialog. For the sake of clarity, we report the (normalized) continuous signals, and not their binary versions (that we used). Precision, recall and accuracy scores related to each individual are also indicated.

all the subjects employed, we reach an average accuracy of 71%, average precision of 67%, and average recall of 40%.

4 Remarks

This work has proposed a gesturing-based approach for performing VAD, the automatic detection of people that speak. The reason for using gestures in VAD, typically performed using speech recordings, is that the use of microphones is difficult or illegal in many scenarios of potential interest, including surveillance of public spaces, monitoring of potentially dangerous plants, etc. The core idea behind the approach is that cognitive sciences have demonstrated that speech and gestures, far from being independent expression modalities, are two faces of the same phenomenon. Therefore, gestures can be considered a reliable evidence of speech taking place at the same time.

The preliminary results presented in this paper provide a quantitative confirmation of the finding above and, most importantly, show that the detection of gesturing activity helps to predict whether a person is speaking or not with an accuracy of 71 percent (on a frame-by-frame basis). While not being conclusive about the possibility of reconstructing the actual turns and of performing diarization, the results are certainly promising in the direction of reconstructing conversational dynamics in absence of audio. This appears particularly important as turn-organization has been widely shown to be fundamental in inferring socially important information such as roles, dominance, personality, etc [71].

Besides, this work shows that it is possible to infer information about missing data (speech in this case) from available evidence (videos in this case). In a surveillance setup like the one of the experiments, this opens two conflicting perspectives: on one hand, surveillance approaches can be significantly improved by predicting

phenomena considered so far non-accessible with the sensors at disposition. On the other hand, privacy protection measures applied so far (i.e., legal limitation on the use of microphones in public spaces) might become obsolete and uneffective. In this respect, experiments of the type presented in this work might change the notion of privacy and of its protection.

Future work can take two major directions: the first is to move from VAD to full diarization. This requires the application of probabilistic sequential models taking into account temporal constraints between neighboring frames and a larger amount of data. The second is to try automatic conversation analysis based on gestures and to verify whether (and to what extent) it is possible to perform tasks like role recognition, conflict detection, etc., typically performed using turn-organization and other conversational cues.

5 Conclusions

The realms of automated surveillance and monitoring tend to focus solely on Computer Vision and Pattern Recognition (CVPR) techniques, neglecting social, affective and emotional aspects of human behavior even if this is, in ultimate analysis, their main subject of interest. Actually, the cross-pollination between social psychology and CVPR could lead to new research questions as well as to application domains that, so far, have not been the subject of attention in the computing community. In this chapter we show how the modeling of groups of people may be performed by considering social and psychological theories: in particular, we analyze the detection of groups, their characterization in terms of social links among the participants, and the inference of speech data from video cues only. Due to our initial good results, we are deeply convinced that the cross-fertilization of human and computer sciences for surveillance and monitoring is going to be inevitably extended, and only in this way a new generation of surveillance systems can be designed, making the necessary jump to go beyond the current technology, so far advanced in incremental steps.

Acknowledgements. The authors will thanks Andrea Fossati, Alessio Del Bue, Loris Bazzani, Alessandro Vinciarelli, Anna Pesarin, Giulia Paggetti, Diego Tosato for their remarkable help in the development of the approaches reported in this chapter.

References

1. Adams, L., Zuckerman, D.: The effects of lighting conditions on personal space requirement. Journal of General Psychology 118(4), 335–340 (1991)
2. Aggarwal, J.K., Park, S.: Human motion: Modeling and recognition of actions and interactions. In: 2nd International Symposium on Proceedings of the 3D Data Processing, Visualization, and Transmission, 3DPVT 2004, pp. 640–647. IEEE Computer Society Press, Washington, DC (2004)
3. Altman, I.: The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding. Brooks/Cole Publishing Company, Monterey, CA (1975)

4. Arulampalam, M., Maskell, S., Gordon, N.: A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. IEEE Transactions on Signal Processing 50, 174–188 (2002)
5. Baxter, J.: Interpersonal spacing in natural settings. Sociometry 33(4), 444–456 (1970)
6. Bazzani, L., Tosato, D., Cristani, M., Farenzena, M., Pagetti, G., Menegaz, G., Murino, V.: Social interactions by visual focus of attention in a three-dimensional environment. Expert. Systems 30(2), 115–127 (2013)
7. Boersma, P.: Accurate short term analysis of the fundamental frequency and the harmonics to noise ratio of a sampled sound. IEEE Transactions on Image Processing 17, 97–110 (1993)
8. Boersma, P.: Praat, a system for doing phonetics by computer. Glot International 5(9/10), 341–345 (2001)
9. Breazeal, C.: Designing Sociable Robots. MIT Press, Cambridge (2002)
10. Brown, L., Tian, Y.: Comparative study of coarse head pose estimation. In: Proc. Motion and Video Computing Workshop, pp. 125–130 (2002)
11. Cassell, J., Steedman, M., Badler, N., Pelachaud, C., Stone, M., Douville, B., Prevost, S., Achorn, B.: Modeling the interaction between speech and gesture. In: Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society, pp. 153–158 (1994)
12. Cheng, Z., Qin, L., Huang, Q., Jiang, S., Tian, Q.: Group activity recognition by gaussian processes estimation. In: 2010 20th International Conference on Pattern Recognition (ICPR), pp. 3228–3231 (August 2010)
13. Cochran, C., Personal, D.: space requirements in indoor versus outdoor locations. Journal of Psychology 117, 121–123 (1984)
14. Cochran, C., Urbanczyk, D., The, S.: The effect of availability of vertical space on personal space. Journal of Psychology 111, 137–140 (1982)
15. Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Bue, A.D., Tosato, D., Menegaz, G., Murino, V.: Social interaction discovery by statistical analysis of f-formations. In: Proceedings of British Machine Vision Conference (2011)
16. Cristani, M., Pesarin, A., Vinciarelli, A., Crocco, M., Murino, V.: Look at who's talking: Voice activity detection by automated gesture analysis. In: Proceedings of the Workshop on Interactive Human Behavior Analysis in Open or Public Spaces, InterHub 2011 (2011)
17. Cristani, M., Murino, V.: Vinciarelli, A.: Socially intelligent surveillance and monitoring: Analysing social dimensions of physical space. In: First IEEE International Workshop on Socially Intelligent Surveillanceand Monitoring (SISM 2010), San Francisco, California (2010)
18. Cristani, M., Paggetti, G., Vinciarelli, A., Bazzani, L., Menegaz, G., Murino, V.: Towards computational proxemics: Inferring social relations from interpersonal distances. In: SocialCom/PASSAT, pp. 290–297 (2011)
19. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. B 39, 1–38 (1977)
20. Duda, R., Hart, P., Stork, D.: Pattern Classification. John Wiley and Sons (2001)
21. Figueiredo, M., Jain, A.: Unsupervised learning of finite mixture models. IEEE Trans. on Pattern Analysis and Machine Intelligence 24(3), 381–396 (2002)
22. Freeman, L.: Social networks and the structure experiment. In: Research Methods in Social Network Analysis, pp. 11–40 (1989)
23. Gall, J., Lempitsky, V.: Class-specific hough forests for object detection. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2009)
24. Gatica-Perez, D.: Automatic nonverbal analysis of social interaction in small groups: a review. Image and Vision Computing (2009)

25. Gatica-Perez, D.: Automatic nonverbal analysis of social interaction in small groups: a review. Image and Vision Computing 27(12), 1775–1787 (2009)
26. Gifford, R., O'Connor, B.: Nonverbal intimacy: clarifying the role of seating distance and orientation. Journal of Nonverbal Behavior 10(4), 207–214 (1986)
27. Griffitt, W., Veitch, R.: Hot and crowded: Influences of population density and temperature on interpersonal affective behavior. Joumal of Personality and Social Psychology 17, 92–98 (1971)
28. Groh, G., Lehmann, A., Reimers, J., Friess, M.R., Schwarz, L.: Detecting social situations from interaction geometry. In: Proceedings of the 2010 IEEE Second International Conference on Social Computing, SOCIALCOM 2010, pp. 1–8. IEEE Computer Society, Washington, DC (2010), http://dx.doi.org/10.1109/SocialCom.2010.11
29. Hall, E.: The hidden dimension. Doubleday New York (1966)
30. Hall, E.: Handbookfor proxemic research. Studies in the anthropologyof visual communication series. Society for the Anthropology of Visual Communication, Washington, DC (1974)
31. Hall, R.: The hidden dimension, New York (1966)
32. Helbing, D., Molnár, P.: Social force model for pedestrian dynamics. Physical Review E 51(5), 4282–4287 (1995)
33. Heshka, S., Nelson, Y.: Interpersonal speaking distance as a function of age, sex, and relationship. Sociometry 35(4), 491–498 (1972)
34. Hongeng, S., Nevatia, R.: Large-scale event detection using semi-hidden markov models. In: IEEE International Conference on Computer Vision, vol. 2 (2003)
35. Hung, H., Ba, S.O.: Speech/non-speech detection in meetings from automatically extracted low resolution visual features. In: ICASSP, pp. 830–833 (2010)
36. Ivanov, Y.A., Bobick, A.F.: Recognition of visual activities and interactions by stochastic parsing. IEEE Trans. Pattern Anal. Mach. Intell. 22, 852–872 (2000)
37. Jebara, T., Pentland, A.: Action reaction learning: Automatic visual analysis and synthesis of interactive behaviour. In: Proceedings of the First International Conference on Computer Vision Systems, ICVS 1999, pp. 273–292. Springer, London (1999)
38. Kendon, A.: Gesticulation and speech: Two aspects of the process of utterance. The Relationship of verbal and Nonverbal Communication, 207–227 (1980)
39. Kendon, A.: Conducting Interaction: Patterns of behavior in focused encounters (1990)
40. Kendon, A.: Language and gesture: unity or duality?, pp. 47–63. Cambridge University Press (2000)
41. Khondaker, A., Ghulam, M.: Improved noise reduction with pitch enabled voice activity detection. In: ISIVC 2008 (2008)
42. Knapp, M., Hall, J.: Nonverbal Communication in Human Interaction. Harcourt Brace College Publishers (1972)
43. Koay, K.L., Syrdal, D.S., Walters, M.L., Dautenhahn, K.: Living with robots: Investigating the habituation effect in participants? preferences during a longitudinal human-robot interaction study. In: ROMAN 2007 the 16th IEEE International Symposium on Robot and Human Interactive Communication, pp. 564–569 (2007), http://hdl.handle.net/2299/1880
44. Kuzuoka, H., Suzuki, Y., Yamashita, J., Yamazaki, K.: Reconfiguring spatial formation arrangement by robot body orientation. In: Proceeding of the 5th ACM/IEEE International Conference on Human-Robot Interaction, HRI 2010, pp. 285–292. ACM, New York (2010), http://doi.acm.org/10.1145/1734454.1734557
45. Lan, T., Wang, Y., Yang, W., Mori, G.: Beyond actions: Discriminative models for contextual group activities. In: Advances in Neural Information Processing Systems, NIPS (2010)

46. Lanz, O.: Approximate bayesian multibody tracking. IEEE Trans. on Pattern Analysis and Machine Intelligence (2006)
47. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: ECCV Workshop on Statistical Learning in Computer Vision, pp. 17–32 (2004)
48. Lin, W., Sun, M.T., Poovendran, R., Zhang, Z.: 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)
49. Lott, D., Sommer, R.: Seating arrangements and status. Journal of Personality and Social Psychology 7(1), 90–95 (1967)
50. Mantel, N.: The detection of disease clustering and a generalized regression approach. Cancer Research 27(2), 209 (1967)
51. Mazur, A.: On Wilson's Sociobiology. American Journal of Sociology 82(3), 697–700 (1976)
52. McNeill, D.: Hand and mind: What gestures reveal about thought. Chicago University Press, Chicago (1992)
53. Michalowski, M.P.: A spatial model of engagement for a social robot. In: Proceedings of the 9th International Workshop on Advanced Motion Control, AMC 2006 (2006)
54. Nakauchi, Y., Simmons, R.: A social robot that stands in line. In: Proceedings of the Conference on Intelligent Robots and Systems (IROS 2000) (October 2000)
55. Ni, B., Yan, S., Kassim, A.A.: Recognizing human group activities with localized causalities. In: CVPR 2009, pp. 1470–1477 (2009)
56. Oliver, N., Rosario, B., Pentland, A.: Graphical models for recognising human interactions. In: Advances in Neural Information Processing Systems (1998)
57. Pacchierotti, E., Christensen, H.I., Jensfelt, P.: Human-robot embodied interaction in hallway settings: A pilot user study. In: Proceedings of the 2005 IEEE International Workshop on Robots and Human Interactive Communication, pp. 164–171 (2005)
58. Park, S., Trivedi, M.M.: Multi-person interaction and activity analysis: a synergistic track- and body-level analysis framework. Mach. Vision Appl. 18, 151–166 (2007)
59. Pellegrini, S., Ess, A., Schindler, K., Gool, L.V.: You'll never walk alone: modeling social behavior for multi-target tracking. In: Proc. 12th International Conference on Computer Vision, Kyoto, Japan (2009)
60. Richmond, V., McCroskey, J.: Nonverbal Behaviors in interpersonal relations. Allyn and Bacon (1995)
61. Robertson, N., Reid, I.D.: Estimating gaze direction from low-resolution faces in video. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 402–415. Springer, Heidelberg (2006)
62. Robertson, N., Reid, I.: Automatic reasoning about causal events in surveillance video 2011 (2011)
63. Russo, N.: Connotation of seating arrangements. The Cornell Journal of Social Relations 2(1), 37–44 (1967)
64. Savinar, J.: The effects of ceiling height on personal space. Man-Environment Systems 5, 321–324 (1975)
65. Scovanner, P., Tappen, M.: Learning pedestrian dynamics from the real world, pp. 381–388 (2009)
66. Smith, H.: Territorial spacing on a beach revisited: A cross-national exploration. Social Psychology Quarterly, 132–137 (1981)
67. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: Int. Conf. Computer Vision and Pattern Recognition (CVPR 1999), vol. 2, pp. 246–252 (1999)

68. Takayama, L., Pantofaru, C.: Influences on proxemic behaviors in human-robot interaction. In: Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009, pp. 5495–5502. IEEE Press, Piscataway (2009), http://portal.acm.org/citation.cfm?id=1732643.1732940
69. Tosato, D., Farenzena, M., Spera, M., Murino, V., Cristani, M.: Multi-class classification on riemannian manifolds for video surveillance. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 378–391. Springer, Heidelberg (2010)
70. Vinciarelli, A., Pantic, M., Bourlard, H.: Social Signal Processing: Survey of an emerging domain. Image and Vision Computing Journal 27(12), 1743–1759 (2009)
71. Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., Schröder, M.: Bridging the gap between social animal and unsocial machine: A survey of social signal processing. IEEE Transactions on Affective Computing (2011) (to appear)
72. Wang, X., Ma, X., Grimson, W.E.L.: Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. IEEE Trans. Pattern Anal. Mach. Intell. 31, 539–555 (2009)
73. Watson, O.: Proxemic behavior: A cross-cultural study. Mouton De Gruyter (1970)
74. Wells, G., Petty, R.: The effects of over head movements on persuasion. Basic and Applied Social Psychology 1(3), 219–230 (1980)
75. White, M.J.: Interpersonal distance as affected by room size, status, and sex. The Journal of Social Psychology 95(2), 241–249 (1975)
76. Zen, G., Lepri, B., Ricci, E., Lanz, O.: Space speaks: towards socially and personality aware visual surveillance. In: Proceedings of the 1st ACM International Workshop on Multimodal Pervasive Video Analysis, MPVA 2010, pp. 37–42. ACM, New York (2010)