Social interaction discovery by statistical analysis of F-formations

Marco Cristani[1,2]
marco.cristani@univr.it

Loris Bazzani[1]
loris.bazzani@univr.it

Giulia Paggetti[1]
giulia.paggetti@univr.it

Andrea Fossati[3]
fossati@vision.ee.ethz.ch

Diego Tosato[1]
diego.tosato@univr.it

Alessio Del Bue[2]
alessio.delbue@iit.it

Gloria Menegaz[1]
gloria.menegaz@univr.it

Vittorio Murino[1,2]
vittorio.murino@iit.it

[1] Dipartimento di Informatica
University of Verona
Italy

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

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

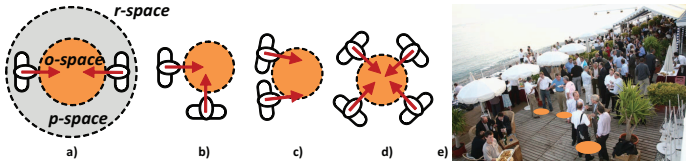

Figure 1: F-formations: a-d) The component spaces of an F-formation: vis-a-vis, L, side-by-side, and circular F-formations, respectively. O-spaces are in orange. e) O-spaces (in orange) in a party scene.

Detecting human interactions represents one of the most intriguing frontiers in the automated surveillance since more than a decade. Very recently, sociologic reasoning has been incorporated into videosurveillance algorithms, and this work follows this way as it attempts to discover social interactions using statistical analysis of spatial-orientational arrangements that have a sociological relevance.

In particular, we analyze quasi-stationary people in an unconstrained scenario identifying those subjects engaged in a face-to-face interaction, i.e., we are dealing with a scene monitored by a single camera where a variable amount of people (10-20) is present; as input, we need the tracking and head orientation data of every individual in the scene. We import into the analysis the sociological concept of F-formation as defined by Adam Kendon in the late '70s [1], commonly adopted in the sociology literature.

More specifically, F-formations are spatial patterns maintained during social interactions by two or more people, and are 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 where every participant looks inward into it, and no external people is allowed in this region (*"no-intrusion"* condition). 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. When there are more than three participants, a circular formation is typically formed. 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 people 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 with 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. Given these (hard) constraints, each k-th subject votes for a candidate center $C(k)$ of the o-space, which has coordinates $x_{C(k)}, y_{C(k)}$ given by by the following voting scheme:

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

where the radius $r = d/2 = 0.75$. Each vote is accumulated in an *intensity accumulation space* $\mathcal{A}_I$, at entry $\tilde{x}_{C(k)}, \tilde{y}_{C(k)}$, where the tilde refers to the closest integer approximation determined by the discretisation of the space $\mathcal{A}_I$. At the same time, the *ID labels* i and j are stored at the same entry of a *label accumulation space* $\mathcal{A}_L$, having the same size of $\mathcal{A}_I$. In the toy example of Fig. 2a, both persons vote for a coincident location (Fig. 2b), which becomes the center of a *candidate* o-space (Fig. 2c). This information is encoded and can be retrieved accessing to the matrices $\mathcal{A}_I$ and $\mathcal{A}_L$, and it is easy to see that if more persons are participating in the interaction they are simply managed by the same voting procedure. At this point, the important condition of *"no-intrusion"* should be checked for the sociological consistence of the candidate o-space. If this condition is verified the candidate o-space finally becomes a *valid* o-space.

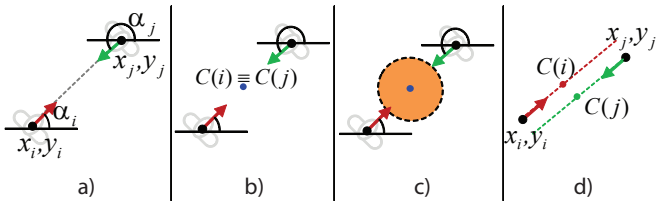

Figure 2: Scheme exemplifying the proposed approach. (a-c) Two subjects exactly facing each other at a fixed distance vote for a same center of the circumference representing the o-space. (d) The 2 subjects do not face each other exactly in real cases.

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. 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. So, in order to deal with real cases, we inject uncertainty in the voting procedure, associating each position and related head orientation to a random variable, following a Gaussian distribution, and applying the same voting method. This process is applied for each person detected in the scene, and finally maximum values of the accumulator spaces $\mathcal{A}_I$ and $\mathcal{A}_L$ provides the most likely interaction group while identifying the interactants too.

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. The outdoor situation is represented by a brand new dataset, dubbed *CoffeBreak* and downloadable, together with the synthetic dataset, at `http://profs.sci.univr.it/~cristanm/datasets.html`. These results are obtained dealing with a complex scenario in which the detection of the persons, the orientation of their heads, and tracking are difficult.

To date, this is the first approach in the literature that discovers social interactions based on the automatic detection of F-formations solely from visual cues. Still, many improvements can be envisaged as future developments.

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