Social Interactions by Visual Focus Of Attention in a three-dimensional environment

M.Farenzena[1], A.Tavano[3], L.Bazzani[1], D.Tosato[1],
G.Paggetti[1], G.Menegaz[1], V.Murino[1,2], M.Cristani[1,2]

[1] Dipartimento di Informatica, Università di Verona, Italy
[2] IIT, Istituto Italiano di Tecnologia, Genova, Italy
[3] University of Leipzig, Leipzig, Germany

Abstract. In the human behavior analysis, the Visual Focus Of Attention (VFOA) of a person is a very important cue. Its detection is difficult, though, especially in a unconstrained and crowded environment, typical of video surveillance scenarios. In this paper, we estimate the VFOA by defining the Subjective View Frustum, which approximates the visual field of a person in a 3D representation of the scene. This opens up to several intriguing behavioral investigations. Here we propose the Inter-Relation Pattern Matrix, that suggests possible social interactions between the people present in the scene. Theoretical justifications and various experimental results substantiate the goodness of the analysis performed.

1 Introduction

Social signal processing aims at developing theories and algorithms that codify how the human behaves while involved in social interactions, putting together perspectives from sociology, psychology and computer science [1–3]. Here the main tools for the analysis are the social signals [3], *i.e.*, temporal co-occurences of social cues [4], that can be basically defined as a set of temporally sequenced changes in neuromuscular, neurocognitive and neurophysiological activity. Social cues are organized into five categories that are heterogeneous, multimodal aspects of a social interplay [3]: 1) *physical appearance*, 2) *gesture and posture*, 3) *face and eyes behavior*, 4) *vocal behavior* and 5) *space and environment*.

In this paper, we concentrate on the Visual Focus Of Attention (VFOA) cue [5–7], that belongs to the third category above, and it is a very important aspect of non verbal communication. The VFOA indicates where and what a person is looking at and is mainly determined by head pose and eye gaze dynamics. In cases where the scale of the scene does not allow to capture the eye gaze directly, though, that can be reasonably approximated by just measuring the head pose [5, 7, 8, 24]. Following this assumption, and considering a general, not restricted scenario, where people can enter, leave, and move freely, we represent VFOA as the *Subjective View Frustum* (SVF), first proposed in [15]. This feature approximates the three-dimensional (3D) visual field of a human subject inside a scene.

According to biological evidence [9], this SVF is modeled as a 3D polyhedron bounding the portion of the scene that the subject is looking at (see Figure 1).

Employing the SVF in conjunction with cues of the *space and environment* category allows to detect social signals of people's interest, with respect to both the physical environment [15], and the other participants acting in the scene. In this paper we propose to statistically infer if a participant is involved in an interactional exchange. In accordance with cognitive and social signalling studies, it is highly likely that interaction takes place when two people are closer than 2 meters [3], and looking at each other [10–12]. We assume that this condition can be reliably inferred by the position and orientation of the SVFs of the people involved. This information can then be gathered in a *Inter-Relation Pattern Matrix* (IRPM), that summarizes the social exchanges occurred between all the participants.

Our proposal is a step forward automatic inference and analysis of social interactions in general, unconstrained, conditions. It represents an alternative with respect to the paradigm of wearable computing [13, 14], or smart rooms [8]. In the typical non-cooperative video surveillance context or when a huge amount of data is required, wearable devices are not usable. Moreover, if a non invasive technology is used, people are more prone to act normally.

Summarizing, this paper provides two novel contributions. First, a more accurate estimation of the Subjective View Frustum: in [15], head orientation is estimated by walking trajectory of the person. This is reasonable when he/she is moving around the scene, but it is not valid in general. We introduce here a more reliable head orientation classification, employing a multi-class boosting algorithm, operating on covariance features [16]. Second, we introduce the Inter-Relation Pattern Matrix, aimed at inferring social interactions among people in a crowded, general scenario.

The rest of the paper is organized as follows: in Sec. 2 the building process of the SVF is explained, sketching all the steps involved. In Sec. 3, the Inter-Relation Pattern Matrix estimation is detailed; then, in Sec. 4 experiments on home-made and public datasets are reported, and, finally, in Sec. 5, conclusions are drawn and future perspectives envisaged.

2 Subjective View Frustum Estimation

Subjective View Frustum (SVF) is defined as the polyhedron $\mathcal{D}$ depicted in Figure 1. It is composed by three planes that delimit the angle of view on the left, right and top sides, in such a way that the angle view is 120° in both directions. The 3D coordinates of the points corresponding to the head and feet of a subject are obtained from the multi-object tracker, while the SVF orientation is obtained by the head pose detector. Our system is therefore composed by four modules, operating in cascade. First, the camera is calibrated and a (rough) 3D model of the scene is constructed. Secondly, a multi-object tracker detects the people position in each frame. Then, this position is used to guide the head pose detector. Finally, all the information is used to estimate the SVF. Each single module is detailed in the following.

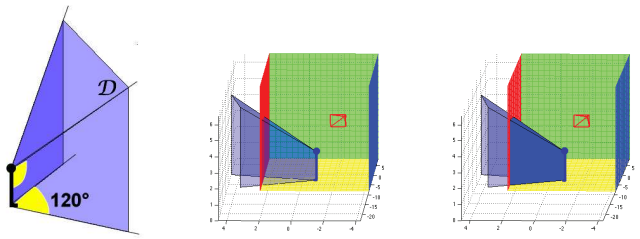


Fig. 1. On the left the SVF model. At the center an example of SVF inside a 3D "box" scene. In red, the surveillance camera position. The SVF orientation is estimated wrt the principal axis of the camera. On the right, in solid blue the same SVF delimited by the scene constraints.

2.1 3D Scene Estimation

In this paper we suppose that the camera monitoring the area is fully calibrated. For convenience, the world reference system is put on the ground plane, with the z-axis pointing upwards. This permits to obtain the 3D coordinates of a point in the image if the elevation from the ground plane is known.

A rough reconstruction of the area, made up of the principal planes present in the scene, can therefore be carried out. An example is shown in Figure 1. This operation requires very little effort. In principle, a more detailed 3D map can be considered, if for example a CAD model of the scene is available or if a Structure-from-Motion algorithm [17, 18] is applied. The choice depends on which level of detail one is willing to gather from the SVF applications.

2.2 Tracking

Our framework needs to cope with several moving entities, prone to severe occlusions that may hold for long time; therefore, a reliable multi-object tracking has to be employed. In this paper we use the Hybrid Joint-Separable (HJS) filter [19], whose characteristics make it well-suited for the task-at-hand. It is essentially a multi-hypothesis particle filtering approach, able to sample in a very efficient way the state space of the system. In practice, a state of the system consists in a joint snapshot of the features that characterize the whereabout of the different moving entities (in our case position, velocity, appearance). During the tracking, HJS decouples and recombines this joint state as a product of single states, each of them encoding a single subject. This permits to keep the computational effort low, that in fact grows linearly with respect to the number of people present in the scene.

2.3 Head Direction Estimation

The tracker provides the location of head and feet for each person in each frame. On the head approximate position, we carry out a multi-class algorithm that recovers the head direction. At the scale of a typical video surveillance scenario,

this estimation is very challenging, since head is depicted very small in the image. Thus, we are content with a rough estimator, able to classify four classes, representing the four possible directions (North, South, East, West) related to the camera orientation. We use a natural multi-class extension of the binary classifier for pedestrian detection presented in [16]. It is based on the estimation of covariance features, i.e. covariance matrix of image characteristics such as spatial location, intensity, higher order derivatives, etc. Classification is realized by a cascade of weak classifiers trained using a multi-class LogitBoost algorithm [20]. We train the classifier with a home-made head dataset, obtained by considering the PETS 2007 dataset[4].

2.4 Subjective View Frustum

The SVF $\mathcal{D}$ is computed precisely using Computational Geometry techniques. It can be written as the intersection of three negative half-spaces defined by their supporting planes respectively of the left, right and top side of the subject. In principle the SVF is not bounded in depth, modeling the human capability of focusing possibly on a remote point located at infinite distance. In practice, though, the SVF is limited by the planes that set up the scene, according to the 3D scene (see Figure 1). The scene volume is similarly modeled as intersection of negative half-spaces. Thus, the exact SVF inside the scene can be computed solving a simple *vertex enumeration* problem, for which very efficient algorithms exist in literature [21].

3 Inter-Relation Pattern Matrix

Subjective View Frustum can be employed as a tool to uncover the visual dynamics of interactional interactions among two or more people. Such an analysis relies on few assumptions with respect to social cues, i.e., that the entities involved in *social interactions* stand closer than 2 meters (covering thus the *socio-consultive zone* – between 1 and 2 meters – the *casual-personal zone* – between 0.5 and 1.2 meters – and the *intimate zone* – around 0.4-0.5 meters), as pointed out in [3]; secondly, it is generally well-accepted that initiators of conversations often wait for visual cues of attention, in particular, the establishment of eye contact, before launching into their conversation during unplanned face-to-face encounters [10–12]; in this sense, SVF may be employed in order to infer whether an eye contact occurs among close subjects. This happens with high probability when the following conditions are satisfied: 1) the subjects are closer than 2 meters; 2) their SVFs overlap and 3) their heads are positioned inside the reciprocal SVFs (see Figure 2). The Inter-Relation Pattern Matrix (IRPM) records when a possible social interaction occurs, and it can be formalized as a three dimensional matrix [22], where each entry $(i, j, t) = (j, i, t)$ is set to one if subjects i and j satisfy the three conditions above, during the $t-$th time interval. As we will see in the experiments, IRPM permits to estimate the number of continuous

[4] `http://pets2007.net/`

interactions and their durations occurring in the scene monitored, which is the basic step for inferring several social signals [23].

Fig. 2. On the left, a frame in which two people are talking to each other. On the right, top view of their SVFs. The estimated orientation, East for 1 and West for 2, is relative to the camera orientation (the pyramid in red in the picture). The SVFs satisfy the three conditions explained in Section 3.

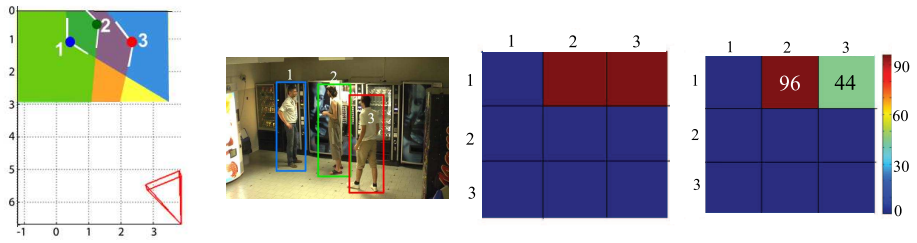
4 Experiments

The experiments aim at showing the usage of the IRPM, and its capability in individuating social exchanges. In specific, we focused on two videosequence datasets. The first one is home-made, and portrays a coffee-room scenario (one frame is depicted in Figure 2), where students take coffee or discuss. The video footage was acquired with an off-the-shelf monocular camera, located on a upper angle of the room. The total amount of video data covered about 30 minutes. The people involved in the experiments were 6, and they were not aware of the scope of the experiment, behaving naturally.

For each frame we build a three-dimensional IRPM, that tells which people are potentially interacting at that moment. In Figure 3 we show one frame example and the SVFs of the people detected. The resulting IRPM is reported in Figure 3c.[5] It must be noticed that by SVFs intersection we capture that person 1 is in relation with 2 and 3. Since 2 and 3 do not look at each other, they are not socializing, according to the definition of social interaction of Section 3. It is worth noting that the sum along t direction gives a summary of what happened during the subsequence (Figure 3d): for the purposes of this analysis, we defined a *social exchange* as a continuative interaction between two people lasting at least 10 seconds.
In these hypothesis, we can infer that there is a social exchange between subjects 1 and 2 and between 1 and 3.

In order to validate our framework, we analyzed all the subsequences and we asked the subjects present in each one if they had any interactions with the other participants, during at least 10 seconds, and with whom. In this way, we label our dataset, obtaining ground-truth data, highlighting 19 social exchanges. Then, we checked if our framework revealed the same dyadic interactions.

[5] Being the IRPMs symmetric and having null main diagonals, we report for clarity only their strictly upper triangular parts.



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

Fig. 3. Analysis of the Inter-relations in one frame of the *Coffee Room* sequence. In (a), the estimated SFVs. The white lines emphasize the SVF position and orientation. In (b), the tracker results, relative to the SVFs in (a). It results that Blue (1) intersects with Green (2) and Red (3). The resulting IR Matrix is (c). In (d) the summary of the subsequence. Each position (i, j) shows the maximum continuative period of time (in seconds) in which the subjects i and j interact (the figures are more explicative if printed in color).

For example, in another subsequence, shown in Figure 4, 5 people enter in the scene and they start talking to each other in circle, the IRPM reveals correctly all but one interaction (4 with 5). This happens because the subject 5 is not correctly tracked during the sequence.

Fig. 4. In the pictures on the left, inter-relations in a group of 5 people, in the *Coffee Room* sequence. On the right, the IR Matrix summarized over t.

Examining all subsequences, and comparing the results with the ground-truth labels, our framework was able to correctly recover about 89% of social exchanges (17/19). There are also 11% false negatives, i.e. the framework reveals a social exchange between people who do not interact. This inaccuracy is mainly due to erroneous tracking and head pose results, particularly challenging when people are grouped together and frequently intersect.

The second database consists of some sequences of the PETS 2007. In this case, the aim is to show the expressiveness of our framework on widely known and used datasets, depicting general, unconstrained scenarios. The sequences taken into account for the experiments are two. They both belong to the $S07$ dataset, in which an airport area is monitored. The first sequence, $S07_{C2}$, is captured by Camera 2, the second one, $S07_{C4}$, is captured by Camera 4, for a total of few minutes of footage. In both cases, validation can not be done as

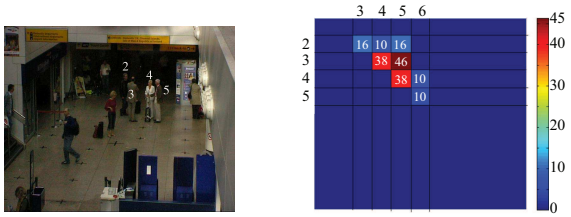


Fig. 5. On the center right, one frame of the sequence. On the right, each filled entry of the colored table shows the sum over t of the related IRPM, considering in the summation only those continuative periods of interactions (> 10 sec., see the text). It suggests that the group in the center of the scene is socially interacting. The interactions with subject 6, not displayed in the frame, are caused by a person crossing the area where the group stays.

above, so we just observe what happens in the sequence and see if the IRPM mirrors the situation occurred in the scene.
The video $S07_{C2}$ shows 4 people in the center of scene, talking to each other. The others just cross the scene or wander on their on, without interacting with anyone. The resulting IR Matrix (Figure 5) depicts this same situation.
The video $S07_{C4}$, instead, shows people passing by, not continuously interacting for a sufficient time interval (10 secs at least) to each other. As expected, the IRPM does not show any social exchanges.

5 Conclusions

In this paper we proposed a novel framework which may help in inferring social signals in a scene. The main feature is the Subjective View Frustum, that encodes the visual field of a person in a 3D environment. The SVF is detected through well-known Computer Vision techniques, and it permits to define novel analysis tools, such as the Inter-Relation Pattern Matrix. We show preliminary but convincing results, that lead to several future improvements: together with a refinement of the head pose detector (in order to find tilt and roll parameters and a more informative pan quantization), it may be possible to analyze also gesture recognition modules, useful to capture different and more complicated social interactions.

Acknowledgements The authors would like to thanks Alessandro Vinciarelli for all the fruitful suggestions and comments.
This research is founded by the EU-Project FP7 SAMURAI, grant FP7-SEC-2007-01 No. 217899.

References

1. Pantic, M., Pentland, A., Nijholt, A.: Special issue on human computing. IEEE SMC, Part B **39**(1) (2009)

2. Pentland, A.: Social signal processing. Signal Processing Magazine, IEEE **24**(4) (2007) 108–111
3. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: Survey of an emerging domain. Image and Vision Computing (2008)
4. Ambady, N., Rosenthal, R.: Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. Psychological bulletin **111**(2) (1992) 256–274
5. Stiefelhagen, R., Finke, M., Yang, J., Waibel, A.: From gaze to focus of attention. In: VISUAL '99, London, UK, SpringerVerlag (1999) 761–768
6. Liu, X., Krahnstoever, N., Ting, Y., Tu, P.: What are customers looking at? In: Advanced Video and Signal Based Surveillance, 2007. (2007) 405–410
7. Smith, K., Ba, S., Odobez, J., Gatica-Perez, D.: Tracking the visual focus of attention for a varying number of wandering people. IEEE PAMI **30**(7) (2008) 1–18
8. Wright, O.R.J.: Summary of research on the selection interview since 1964. In: MLMI06. (2006)
9. Panero, J., Zelnik, M.: Human Dimension and Interior Space : A Source Book of Design Reference Standards. Whitney Library of Design, New York (1979)
10. Langton, S.H.R., Watt, R.J, Bruce, V.: Do the eyes have it? cues to the direction of social attention. Trends in Cognitive Neuroscience **4**(2) (2000) 50–58
11. Whittaker, S., Frohlich, D., Daly-Jones, O.: Informal workplace communication: what is it like and how might we support it? In: CHI '94, New York, NY, USA, ACM (1994) 208
12. Jabarin, B., Wu, J., Vertegaal, R., Grigorov, L.: Establishing remote conversations through eye contact with physical awareness proxies. In: CHI '03 extended abstracts, New York, NY, USA, ACM (2003)
13. Pentland, A.: Looking at people: Sensing for ubiquitous and wearable computing. IEEE PAMI. **22**(1) (2000) 107–119
14. Choudhury, T., Pentland, A.: The sociometer: A wearable device for understanding human networks. In: CSCW - Workshop on ACCUCE. (2002)
15. Farenzena, M., Bazzani, L., Murino, V., Cristani, M.: Towards a subject-centered analysis for automated video surveillance. In: ICIAP. (2009)
16. Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on riemann manifold. IEEE PAMI **30**(10) 1713–1727
17. Farenzena, M., Fusiello, A., Gherardi, R., Toldo, R.: Towards unsupervised reconstruction of architectural models. In: Proceedings of VMV 2008. (2008) 41–50
18. Snavely, N., Seitz, S., Szeliski, R.: Photo tourism: exploring photo collections in 3D. In: SIGGRAPH, NY, USA (2006) 835–846
19. Lanz, O.: Approximate bayesian multibody tracking. IEEE PAMI **28**(9) (2006) 1436–1449
20. Zhu, J., Rosset, S., Zou, H., Hastie, T.: Multiclass adaboost. Technical report, Stanford University (2005)
21. Preparata, F.P., Shamos, M.I. Computational Geometry. An Introduction. Springer-Verlag (1985) 72–77
22. Freeman, L.: Social networks and the structure experiment. In: Research Methods in Social Network Analysis (1989) 11–40
23. Gatica-Perez, D.: Automatic nonverbal analysis of social interaction in small groups: a review. Image and Vision Computing (2009) in press
24. Lanz, O., Brunelli, R., Chippendale, P.I., Voit, M., and Stiefelhagen, R.: Extracting Interaction Cues: Focus of Attention, Body Pose, and Gestures. In: Computers in the Human Interaction Loop. Springer (2009) 87–93