PRAI*HBA special issue: Social Interactions by Visual Focus of Attention in a Three-Dimensional Environment

L. Bazzani[a], M. Cristani[a,b], D. Tosato[a], M. Farenzena[a], G. Paggetti[a], G. Menegaz[a], V. Murino[a,b]

[a] *Dipartimento di Informatica, Università di Verona, Italy*
[b] *IIT, Istituto Italiano di Tecnologia, Genova, Italy*

Abstract

In human behavior analysis, the Visual Focus Of Attention (VFOA) of a person is a very important cue. VFOA 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. In particular, we propose the Inter-Relation Pattern Matrix, that suggests possible social interactions between the people present in a scene. Theoretical justifications and experimental results substantiate the validity and the goodness of the analysis performed.

Keywords: Social Signaling, Visual Focus of Attention, Social Interactions, Tracking, Head Pose Estimation

1. Introduction

The automatic recognition of human activities in video recordings is undoubtedly one of the main challenges for a surveillance system. This is usually accomplished using a serial architecture built upon an array of techniques aimed at extracting low-level information including, for instance, foreground/background segmentation (Benezeth et al., 2008) and object tracking (Fuentes and Velastin, 2006). After these early processing stages, high-level analysis methods aim at detecting atomic actions (e.g., gestures) as well as complex activities (i.e., spatio-temporal structures composed of atomic actions) (Chellappa et al., 2005), possibly exploiting ontologies for ensuring interoperability across different platforms and semantic descriptions understandable to human operators (Francois et al., 2005).

However, these technologies seem to forget that, for human beings, physical and social space are tightly intertwined and no intelligent monitoring is possible without taking into account social aspects associated to behaviors displayed under the eyes of the cameras. This is especially regrettable when other domains, e.g. Affective Computing (AC) (Picard, 2000) or Social Signal Processing (SSP) (Vinciarelli et al., 2009), pay significant attention to social, affective and emotional aspects of human behavior.

In particular, Social Signal Processing aims at developing theories and algorithms that codify how human beings behave while involved in social interactions, putting together perspectives from sociology, psychology, and computer science (Pentland, 2007; Vinciarelli et al., 2009; Pantic et al., 2009). Here, the main tools for the analysis are the social signals (Vinciarelli et al., 2009), *i.e.*, temporal co-occurences of social cues (Ambady and Rosenthal, 1992), 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 (Vinciarelli et al., 2009): 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 (Stiefelhagen et al., 1999; Liu et al., 2007; Smith et al., 2008), that belongs to the third category, and it is a very important aspect of non verbal communication; as explained later, we take also into account the fifth category, usually disregarded by social signaling studies (Cristani et al., 2010). The VFOA indicates where and what a person is looking at and it is mainly determined by head pose and eye gaze estimation. In cases where the scale of the scene does not allow to capture the eye gaze directly, viewing direction can be reasonably approximated by just measuring the head pose; this assumption has been exploited in several approaches dealing with a meeting scenario (Stiefelhagen et al., 1999, 2002; Voit and Stiefelhagen, 2008) or in a smart environment (Smith et al., 2008; Lanz et al., 2009).

Following this claim, and considering a general, unrestricted scenario, where people can enter, leave, and move freely, we approximate VFOA as the *Subjective View Frustum* (SVF), first proposed in (Farenzena et al., 2009a). This feature represents the three-dimensional (3D) visual field of a human subject in the scene. According to biological evidence (Panero and Zelnik, 1979), the SVF can be modeled as a 3D polyhedron delimiting 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 signals of the possible people's interest, with respect to both the physical environment (Farenzena et al., 2009a), and the other participants acting in the scene. More specifically, we propose a method to statistically infer if a participant is involved in an interactional exchange. In accordance with cognitive and social signaling studies, it is highly probable that the interaction takes place when two persons are closer than 2 meters (Vinciarelli et al., 2009), and looking at each other (Whittaker et al., 1994; Langton et al., 2000; Jabarin et al., 2003). 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 encodes the social exchanges occurred between all the participants.

Detecting social relations among people may be useful to instantiate a more robust definition of group in surveillance applications. Actually, in the last few years, several applications focused on the group modeling have been proposed (Mckenna et al., 2000; Marques et al., 2003) and re-identification (Zheng et al., 2009); in the first application a group is defined following physically-driven proximity principles, while in the re-identification groups are assumed to be detected from an external algorithm.

More in general, our proposal is a step forward automatic inference and analysis of social interactions in general, unconstrained conditions: it is alternative to the paradigm of wearable computing (Pentland, 2000; Choudhury and Pentland, 2002), or smart rooms (Waibel et al., 2003). In the typical non-cooperative video surveillance context or when a huge amount of data is required, wearable devices are not usable. Moreover, the use of non-invasive technology makes people more prone to act normally.

Considering the literature (except our first work in (Farenzena et al., 2009a)), the "subjective" point of view for automated surveillance systems has been taken into account in (Benfold and Reid, 2009), that takes ideas from (Robertson and Reid, 2006), and represents therefore the most similar approach in the literature to ours. In that paper, the goal is to address the head orientation of low-resolution pedestrians to infer interest regions in the scene. The difference with respect to our system is that in their case the gaze orientation was modeled in a continuous way, while we restrict to a fixed number of orientations (=4); in addition, in (Benfold and Reid, 2009), interaction analysis was absent, and the subjective point of view was functional solely on the estimation of interest maps of the scene. This last point is the most important, distinctive aspect of our work.

The works of (Otsuka et al., 2006) and (Hung et al., 2008) are also close to ours as they estimate a sort of focus of attention of single individuals. They are also different from our work since they consider a meeting scenario that is usually more constrained than a surveillance scenario, and that can be monitored with higher accuracy. In (Otsuka et al., 2006), the gaze pose in high-resolution images is estimated to infer inter-personal relations. As mentioned later, we prefer to perform head pose estimation because eye gaze is very hard because of the low resolution. This idea is also followed by (Hung et al., 2008). However, they suppose that the VFOA of each person is constrained: a person can look only at another person. This assumption could be invalid in a surveillance scenario, where people can wander around freely, look at other objects in the environment, be distracted by external events during a conversation in a group and so on. For this reasons, we left unconstrained the head pose estimation.

Summarizing, this paper provides two novel contributions. First, we propose a more accurate estimation of the Subjective View Frustum: in (Farenzena et al., 2009a), head orientation is estimated by walking trajectory of the person. This is reasonable when he/she is moving in 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 (Tuzel et al., 2008). Second, we introduce the Inter-Relation Pattern Matrix, aimed at inferring social interactions among people in a crowded, general scenario. This work not only fills a gap in the state of the art of SSP aimed at understanding social interactions, but also represents a novel research opportunity, alternative to the scenarios considered so far in socially-aware technologies, where automatic analysis techniques for the spatial organization of social encounters are taken into account.

The rest of the paper is organized as follows. In Sec. 2, the main techniques for estimating the VFOA in absence of gaze information and the methods for head pose estimation are reviewed. In Sec. 3, the building process of our SVF estimation method is described, sketching all the involved processing steps. In Sec. 4, the Inter-Relation Pattern Matrix description is reported. Therefore, in Sec. 5, experiments on home-made and public datasets are illustrated, and, finally, in Sec. 6, conclusions are drawn together with possible future developments of the work.

2. State of the art

It is well known that a person's VFOA is determined by his eye gaze. Since objects are foveated for visual acuity, gaze direction generally provides more precise information than other bodily cues regarding the spatial localization of ones attentional focus. A detailed overview of gaze-based VFOA detection in meeting scenarios is presented in (Ba and marc Odobez, 2006). However, measuring the VFOA by using eye gaze is often difficult or impossible: either the movement of the subject is constrained or high-resolution images of the eyes are required, which may not be practical (Matsumoto et al., 2002; Smith et al., 2003), and several approximations are considered in many cases. For example, in (Stiefelhagen et al., 1999), it is claimed that the VFOA can be reasonably inferred by head pose in many cases. Following the same assumption, in (Smith et al., 2008) pan and tilt parameters of the head are estimated, and the VFOA is represented as a vector normal to the person's face. It is employed to infer whether a walking person is focused on an advertisement located on a vertical glass or not. Since the situation is very constrained, this proposed VFOA model works pretty well; anyway, as observed by the authors themselves, a more complex model, that considers camera position, people's position and scene structure, is required in a more general situation. The same considerations hold for the work presented in (Liu et al., 2007), where Active Appearance Models are fit on the face of the person in order to discover which portion of a mall-shelf is observed.

In (Lablack and Djeraba, 2008), the visual field is modeled as a tetrahedron associated with a head pose detector. However, their model fixes the depth of the visual field, and this is quite unrealistic. Our SVF models the visual field as well, but in our case, owing to the 3D environment in which the SVF lives, we let the SVF be bounded by the structure of the scene, which is more reasonable. Moreover, our formulation is not restricted to controlled environments, but it can be employed to analyze any generic scene.

Our proposal extends the work done in Farenzena et al. (2009b), which is the first promoting the use of the visual focus of attention for interaction modeling in a Computer Vision context.

Head Pose Estimation. Head orientation estimation is an important Computer Vision application. Numerous and different are the approaches present in the literature; a recent review has been proposed by (Murphy-Chutorian and Trivedi, 2009), where a performance analysis of different methods are presented, and a list of the commonly used dataset for head pose estimation is shown. Moreover, CLEAR workshops are important events for the head pose estimation community, and several important approaches can be found in the related proceedings (Stiefelhagen and Garofolo, 2007; Stiefelhagen et al., 2008). The main differences between these technique and the proposed head pose estimation method are: 1) Most of the them are multi-view, whereas our approach works also with a single image. 2) The training set we used has very low-resolution images (20 × 20); CLEAR dataset contains 60 × 60 images. In such low resolution, we had to cast the classification problem to few classes (without continuous pose values). 3) Most of the methods in CLEAR proceedings perform also head pose tracking. Our method can be also used with still images.

In the multi-faceted realm of the classification approaches, boosting-based techniques play a primary role (Li et al., 2002; Viola and Jones, 2001; Huang et al., 2005; Wu et al., 2004; Li and Zhang, 2004; Tuzel et al., 2008; Wu and Nevatia, 2008; Paisitkriangkrai et al., 2008). Boosting (Freund and Schapire, 1997; Schapire and Singer, 1999; Friedman et al., 2000) is a remarkable, highly customizable way to create strong and fast classifiers, employing various features fed into diverse architectures with ad-hoc policies. Among the different features exploited for boosting in surveillance applications (see (Wu and Nevatia, 2009) for an updated list), covariance features (Tuzel et al., 2006) have been exploited as powerful descriptors of pedestrians (Tuzel et al., 2008; Wu and Nevatia, 2008), and their effectiveness has been explicitly investigated in a comparative study (Paisitkriangkrai et al., 2008). When injected in boosting systems (Tuzel et al., 2008; Wu and Nevatia, 2008; Paisitkriangkrai et al., 2008), covariances provide strong detection performances, encapsulating possible high intra-class variances (due to pose and view changes of an object of interest). They are in general stable under noise, and furnish an elegant way to fuse multiple low-level features as, in fact, they intrinsically exploit possible inter-feature dependencies.

In (Tuzel et al., 2008), the use of covariance matrix descriptors is tailored for pedestrian detection. In the learning step, given a set of pedestrian and background images, LogitBoost was used for both a greedy estimation of a set of image patches that generate most discriminative covariances matrices among a set of different sizes and positions, and for classifying the images patches themselves, i.e., as feature selection and classification method at the same time. The same reasoning, i.e., using boosting for feature selection and classification, has been applied to other approaches in the literature, as for example in (Wu and Nevatia, 2009; Chen et al., 2009). Recently, the use of regression trees (Breiman et al., 1984) as weak classifiers has been promoted, showing great performances with strong noise.

3. Subjective View Frustum Estimation

The *Subjective View Frustum* (SVF) is defined as the polyhedron $\mathcal{D}$ depicted in Figure 1. It is composed by three planes that delimit the angles of view on the left, right and top sides, in such a way that the angle span is 120° in both directions. The 3D coordinates of the points corresponding to the head and feet of a subject are obtained from a multi-target tracker, while the SVF orientation is obtained by an head pose detector (see below). Our system is therefore composed by four modules operating in cascade.

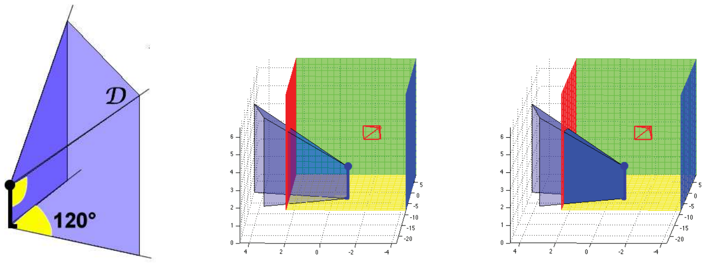


Figure 1: Left: the SVF model. Center: an example of SVF inside a 3D "box" scene. In red, the surveillance camera position: the SVF orientation is estimated with respect to the principal axes of the camera. Right: the same SVF delimited by the scene constraints (in solid blue).

First, the camera is calibrated and a (rough) 3D model of the scene is constructed. Second, a multi-target tracker detects the people position in each frame, and this data 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.

3.1. 3D Scene Estimation

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.

Therefore, a rough reconstruction of the area, made up of the principal planes present in the scene, can be carried out, see an example 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 (Farenzena et al., 2008; Snavely et al., 2006) is applied. The choice depends on which level of detail one is willing to gather from the SVF applications.

3.2. Tracking

Multi-target tracking has been well investigated in literature. In this work, we use a well-known method called Hybrid Joint-Separable (HJS) filter (Lanz, 2006), because it deals with severe occlusions. It is essentially a multi-hypothesis particle filtering approach, able to sample in an efficient way the joint state space of the targets.

From a Bayesian perspective, the single object tracking problem aims at recursively calculating the posterior distribution $p(x_t|z_{1:t})$, where x_t is the current state of the target (*e.g.*, its position), z_t is the current measurement or observation (*e.g.*, the current frame), and $x_{1:t}$ and $z_{1:t}$ are the states and the measurements up to time t, respectively:

$$p(x_t|z_{1:t}) \propto p(z_t|x_t) \int p(x_t|x_{t-1})p(x_{t-1}|z_{1:t-1})dx_{t-1}. \tag{1}$$

This recursive formulation is fully specified by an initial distribution $p(x_0)$, the dynamical model $p(x_t|x_{t-1})$, and the observation model $p(z_t|x_t)$. Particle filtering approximates the posterior distribution by a set of N weighted particles, *i.e.*, $\{(x_t^{(n)}, w_t^{(n)})\}_{n=1}^N$; a large weight $w_t^{(n)} \propto p(z_t|x_t^{(n)})$ mirrors a state $x_t^{(n)}$ with high posterior probability. Thus, particle filtering consists in generating new hypothesis according to $p(x_t|x_{t-1})$ and evaluating their likelihood $p(z_t|x_t)$.

HJS filter is an extension of this framework for multiple targets. It adopts the approximation $p(\mathbf{x}_t|z_{1:t}) \approx \prod_k p(x_t^k|z_{1:t})$, that is, the joint posterior $\mathbf{x}_t = \{x_t^1, x_t^2, \ldots, x_t^K\}$ could be approximated via the product of its marginal components (k indexes the individual targets). The dynamics and the observation models of HJS are marginalized out as follows:

$$p(x_t^k|x_{t-1}^k) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}^{\neg k}|z_{1:t-1})d\mathbf{x}_{t-1:t}^{\neg k} \tag{2}$$

$$p(z_t|x_t^k) = \int p(z_t|\mathbf{x}_t)p(\mathbf{x}_t^{\neg k}|z_{1:t-1})d\mathbf{x}_t^{\neg k} \tag{3}$$

where $^{\neg k}$ means all the targets but the kth. These equations encode an intuitive strategy: the dynamics and the observation models of the kth target lie upon the consideration of a joint dynamical model $p(\mathbf{x}_t|\mathbf{x}_{t-1}) \approx p(\mathbf{x}_t) \prod_k q(x_t^k|x_{t-1}^k)$ and $p(z_t|\mathbf{x}_t)$, respectivelly. The joint distribution $p(\mathbf{x}_t)$ avoids that multiple targets with single motion $q(x_t^k|x_{t-1}^k)$ collapse in a single location. Please note that $p(x_t^k|x_{t-1}^k)$ is different from $q(x_t^k|x_{t-1}^k)$, since that $q(x_t^k|x_{t-1}^k)$ does not take into account the interactions between targets, whereas $p(x_t^k|x_{t-1}^k)$ does it because it is integrated over $\mathbf{x}_{t-1:t}^{\neg k}$. The joint observation model considers that the visual appearance of a single target may be occluded by another object simulating a z-buffer. The two models are weighted by posterior distributions that essentially promote trusted joint objects configurations (not considering the kth object). For more details about how to compute Eq. 1, 2 and 3, the HJS algorithm and the features used for tracking, readers may refer to the original paper (Lanz, 2006).

3.3. Head Orientation Estimation

The tracker provides the location of the head and the feet for each person in each frame. As for the head approximate position, we define a square window I of size $r \times r$, where we run a multi-class algorithm that recovers the head orientation. The size r has been chosen large enough in order to contain a head, considering the experimental physical environment and the camera position.

For the multi-class classification, we boost regression trees (Breiman et al., 1984), because they are the ideal weak learning strategy, since they can tolerate a significant amount of labeling noise and errors in the training data (which are very likely in low resolution images). Moreover, they are very efficient at runtime, since matching a sample against a tree is logarithmic in the number of leaves.

From the mathematical point of view, they are an alternative approach to nonlinear regression. The principle is to sub-divide, or partition, the space in two smaller regions, where the data distribution is more manageable. This partitioning proceeds recursively, as in hierarchical clustering, until the space is so tame that a simple model can be easily fitted. The global model thus has two parts: one is just the recursive partition, the other is a simple model for each cell of the partition. Regression trees are more powerful than global models, like linear or polynomial regression, where a single predictive formula is supposed to hold over the entire data space.

In order to avoid the risk of overtraining of the regression tree, we establish as stopping rule a minimal number τ of observations per tree leaf, experimentally estimated (see Sec. 5).

In our approach, we extract from each image I ($r \times r$ pixels), a set $\Phi(I, x, y)$ of dimension $r \times r \times d$ features where $d = 12$ and x, y are the pixel locations. It is composed by:

$$\Phi(I, x, y) = \left[X \; Y \; R \; G \; B \; I_x \; I_y \; O \; \mathrm{Gab}_{\{0,\pi/3,\pi/6,4\pi/3\}}\right]. \tag{4}$$

X, Y represent the spatial layout maps in I, and R, G, B are the color channels. I_x and I_y are the directional derivatives of I, and O is the gradient orientation. Finally, Gab is a set of 4 maps containing the results of Gabor filtering, with filters of dimension 2×4, sinusoidal frequency 16, and directions $\mathcal{D} = \{0, \pi/3, \pi/6, 4\pi/3\}$. In order to increase the robustness to local illumination variations, we apply the normalization operator introduced in (Tuzel et al., 2008) before applying the multi-class framework. First, we estimate the covariance of the image I, denoted as X_I. Then, for each element X_i of the dataset, we apply the following normalization:

$$\widehat{X}_i = \mathrm{diag}(X_I)^{-\frac{1}{2}} X_i \, \mathrm{diag}(X_I)^{-\frac{1}{2}}, \tag{5}$$

where $\widehat{X}_i$ is the normalized descriptor, and $\mathrm{diag}(X_I)$ is a square matrix with only the diagonal entries of X_I.

Our approach takes inspiration from the literature on dense image descriptors (see (Dalal and Triggs, 2005) as an example). We sample the window I employing an array of $N_P = 16$ uniformly distributed and overlapping patches of the same dimension. For each sampled patch locations inside the $r \times r$ region of interest, described by the covariance matrix of a set of d image features described by the Eq. (4), a multi-class LogitBoost classifier is trained. Each class represent a different head orientation sampled according with a fixed sampling step α and from an extra class containing all the background examples. We experimentally found that $\alpha = 90°$ which correspond to the semantic classes North, South, East and West, is enough for our purposes. At testing time, each patch of a sample window (Fig. 2) is independently classified. Then, the classification result is given by a majority criterion across the patches. We name the combination of this patch description that encodes the local shape and appearance and its uniformly distributed architecture *ARray of COvariances* (ARCO, for the sake of brevity).

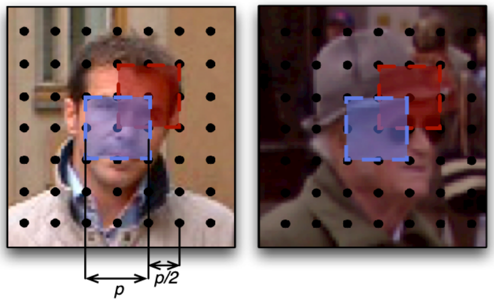


Figure 2: Array of Covariance matrices (ARCO) feature. The image is organized as a grid of uniformly spaced and overlapping patches. The head orientation result of each patch is estimated by a multi-class classifier.

More formally, given a set of patches $\{P_i\}_{i=1,\ldots,N_P}$, we learn a multi-class classifier for each patch location $\{F_{P_i}\}_{i=1,\ldots,N_P}$ through the multi-class LogitBoost algorithm (Friedman et al., 2000), adapted to work on Riemannian manifolds, as suggested by (Tuzel et al., 2008), that basically means each covariance matrix must be projected on a proper tangent space (vector space) of the Riemannian manifold to be classified. Since we deal with a multi-class classification problem, a common tangent space is chosen where all the covariances are projected. For computational convenience, the projection point is the identity matrix I_d. Considering the projection form the mathematical point of view, it is a logarithmic transformation of the (positive) eigenvalues of a covariance matrix. Therefore, the computational complexity of each projection is bounded by the eigenvalue decomposition complexity $O(d^3)$. By the fact that d (we recall that d is the number of image features) is small the projection is a fast operation. All the details of the projection operation are contained in (Tuzel et al., 2008).

Let $\Delta_j = \sum_{i=1}^{N_P}(F_{P_i} == j)$ be the number of patches that vote for the class $j \in \{1,\ldots,J\}$. To assign a class label c to a new image, we fuse the votes with a majority voting strategy among all the classes:

$$c = \arg\max_j \{\Delta_j\}, \quad j = 1,\ldots,J. \tag{6}$$

Actually, in our approach, we employ 5 classes named above, i.e., North, South, East, West, and Background. The first four classes indicate the four directions related to the camera orientation. The Background class is introduced to manage the cases where the tracker fails in providing a correct head position. We are aware that the use of only four directions may lead to rough estimates, but it should be considered that the resolution of the source video data is very poor.

The ARCO representation has several advantages. First, it allows to take into account different features, inheriting their expressivity and exploiting possible correlations, by the use of the covariance local descriptor. In this sense, it could be thought as a compact and powerful integration of features. Second, due to the use of integral images exploited in the computation of the covariance matrices (Tuzel et al., 2008), ARCO is fast to compute, making it suitable for a possible real-time usage.

3.4. 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 of the left, right and top sides of the subject, respectively. In principle, the SVF is not bounded in depth, modeling the human capability of focusing possibly on a remote point located at infinite distance. However, in practice, 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. Consequently, the exact SVF inside the scene can be computed solving a simple *vertex enumeration* problem, for which very efficient algorithms exist in literature (Preparata and Shamos, 1985).

4. The Inter-Relation Pattern Matrix

The SVF can be employed as a tool to discover the visual dynamics of the 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 the *social interaction* 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) (Vinciarelli et al., 2009). Then, 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 (Whittaker et al., 1994; Langton et al., 2000; Jabarin et al., 2003). In this sense, SVF may be employed in order to infer whether an eye contact occurs among close subjects or not. 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 3). In the figure, a 2D projection of the 3D frusta is shown for illustrative purposes. Anyway, the real intersection is calculated between the genuine 3D SVFs. The Inter-Relation Pattern Matrix (IRPM) records when a possible social interaction occurs, and it can be formalized as a three-dimensional matrix (Freeman, 1989), 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 instant.

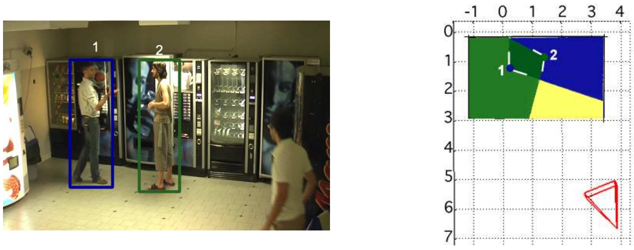


Figure 3: Left: two people are talking each other. Right: top view projection of their SVFs: the estimated orientation, East for 1 and West for 2, is relative to the camera orientation (the pyramid in red in the picture). The SVFs satisfy the three conditions explained in Section 4.

The IRPM matrix serves to analyze time intervals in which we look for social interactions. Let us suppose to focus on the time interval $[t - T + 1, t]$. In this case we take into account all the IRPM slices that fall in $[t - T + 1, t]$, summing them along the t direction, and obtaining the *condensed* IRPM (cIRPM). Intuitively, the higher is the entry $cIRPM_t(i, j)$, the stronger is the probability that subjects i and j are related during the interval $[t-T+1, t]$. Therefore, in order to detect a relation between a pair of individuals i, j in the interval $[t-T+1, t]$, we check if $cIRPM_t(i, j) > Th$, where Th is a threshold a priori defined. This threshold filters out noisy interaction detection: actually, due to the errors in the tracking and in the head pose estimation, the lower the threshold, the higher the possibility of false positive detections. In the experiments, we show how the choice of the parameters T and Th modifies the goodness of the results, in term of social interaction detections.

The cIRPM represents one-to-one exchanges only, but we would like also to capture if there are *groups* in the scene. Here, we will not use the term group in its sociological meaning, but in its common definition. In sociology, a group is usually defined as "a collection consisting of a number of people who share certain aspects, interact with one another, accept rights and obligations as members of the group and share a common identity". We are conscious that the proposed algorithm is not able to identify such complex relations. For this reasons we consider to be correct the use of the common meaning of the term group, that is "an assemblage of objects standing near together, and forming a collective unity; a knot (of people), a cluster (of things)". The latter significance is closer to our aims.

Operationally, we treat the $cIRPM$ as the adjacency matrix of a graph, with a vertex v_i for each people in the scene, and an edge e_{ij} if $cIRPM_t(i,j) > Th$. The *groups* present in the scene are detected by computing the connected components of the graph. Some illustrative examples are depicted in Figures 8, 9 and 10.

5. Experimental Results

The experiments aim at showing the capabilities of the proposed approach. First, we validate the performance of tracking and head orientation classification separately, in order to check the behavior of the single modules. Then, we show how these modules grouped together perform, by analyzing the employment of the IRPM, and its capability in individuating social exchanges.

The evaluation of the multi-target tracker is performed on a publicly available and challenging dataset built for automatic video surveillance purposes: PETS 2009[1]. We carry out a comparative analysis between HJS filter and a state-of-the-art Kalman-based tracker, i.e., Multi-Hypothesis Tracker (MHT) (Blackman, 2004), on a manually annotated sequence. The evaluation proposed by (Smith et al., 2005) is used here, in terms of False Positives (FP), Multiple Objects (MO), False Negatives (FN), Multiple Trackers (MT), and Tracking Success Rate (TSR). Table 1 shows that HJS filter outperforms MHT considering FP, MO, MT, and TSR, because MHT generates more than one tracks for a single target. However, the FN ratio is higher for HJS filter because the tracking of a target could be lost due to occlusions and it converges toward another target or to a clutter observation. TSR that gives us a general value of the tracking reliability and it summarizes the overall performances suggests that HJS filter is better. Qualitative results (Fig. 4) gives the same evidence.

[1] `http://www.cvg.rdg.ac.uk/PETS2009/`

	FP	MO	FN	MT	TSR
MHT	0.279	0.009	**0.203**	0.212	0.624
HJS	**0.086**	**0.007**	0.279	**0.042**	**0.712**

Table 1: Tracking results comparison on PETS 2009 dataset: sequence S2, video L1, view 1.

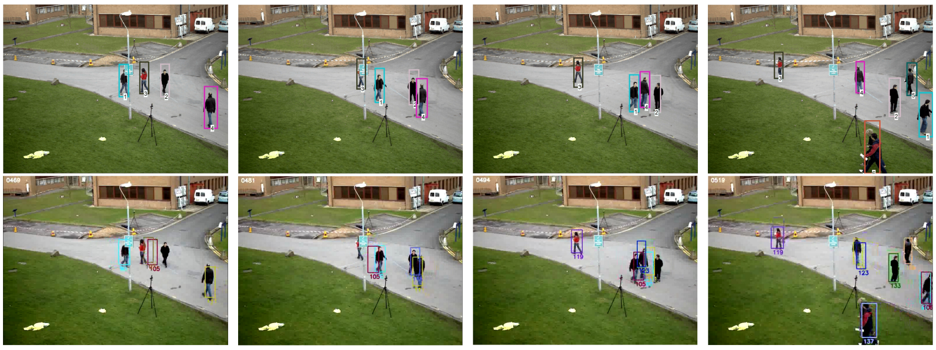


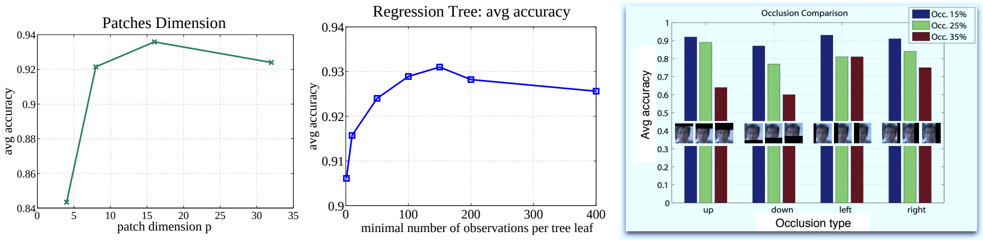
Figure 4: Tracking results comparison on PETS 2009 dataset: sequence S2, video L1, view 1, frames 469, 481, 494 and 519. First row shows MHT and second row show HJS filter.

As to the head orientation classification model, we build a multi-class classifier for head pose classification on the 4 Pose Head Database originally proposed by Orozco et al. (2009) and available at `http://sites.google.com/site/diegotosato/ARCO`. This dataset contains head images of dimension 50 × 50 (see some samples in Fig. 6) obtained from the i-LIDS dataset[2]. These images come from a real video surveillance scene, mirroring well typical critical conditions: they are noisy, motion-blurred, and at low resolution. The images are divided in 4 foreground (FG) classes: Back (4200 examples), Front (3555 examples), Left (3042 examples), and Right (4554 examples). Moreover, this dataset contains another set of 2216 background (BG) images. We partition the dataset in 2 equal parts, using one partition for training and one for testing. For our purposes, we enrich the *original dataset* Orozco et al. (2009) using images ($\sim$ 200 images for each class) coming from our sequence in order to make more robust the final classifier. This *enriched dataset* is available at `http://sites.google.com/site/diegotosato/gdet`. For validation purposes,

[2] `http://scienceandresearch.homeoffice.gov.uk/hosdb/cctv-imaging-technology/i-lids/`

the new images have been equally partitioned into 2 sets, one for the training and one for testing. And all the images have been normalized to a very low-resolution size (20 × 20 is the average resolution of the head images in our videos).

During the training phase, for each image patch, a 5-class classifier is built, as described in Sec. 3.3. Then, give a testing image, the features are extracted and the covariance matrices calculated from all the patches of $p \times p$ pixels, on a fixed grid of $p/2$ pixels steps. This means that the patches remain overlapped by half of their size. In Figure 5(a), we vary p in order to investigate how the dimension of the patches modifies the classification performances. The best performance is obtained with $p = 0.32r$, where $r \times r$ is the image dimension. The τ parameter, that rules the complexity of the regression trees, has been fixed to the optimal value 150 according to the accuracy test in Fig. 5(b). It is interesting to note that exceeding this value, the performance drops, which is a sign of overtraining of the system. Moreover, we test the ability of our classifier to deal with occlusions. Indeed, patch-based classifiers, as part-based classifiers, are naturally able to manage the presence of occlusions. We depict in Figure 5(c) the robustness to four types of occlusions (left-, right-, top- and bottom-side), in different sizes.



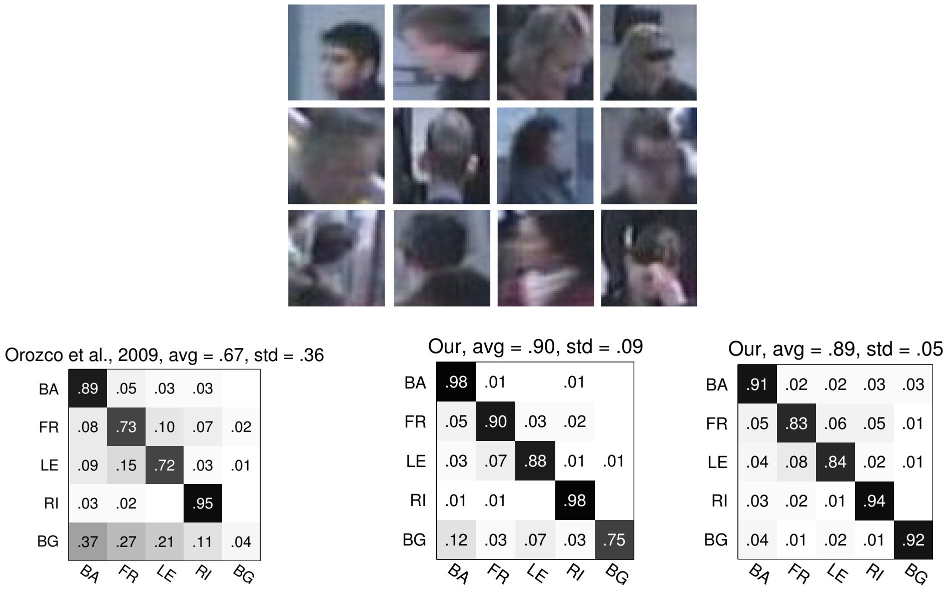
(a)

(b)

(c)

Figure 5: Classification performance on the 4 Head original dataset in terms of mean classification accuracy varying (a) the patch size p, (b) the regression tree stopping criterion (the number of elements per leaf τ) and (c) considering occlusions of different strength.

We compare our method with Orozco et at. (Orozco et al., 2009), the state-of-the-art method for head pose classification for low resolution data. In this work, it is proposed a head pose descriptor based on similarity distance maps to mean appearance templates of head images at different poses. All images in this dataset have their related pose descriptors, provided by the authors themselves (Orozco et al., 2009). The classifier is trained by Support Vector Machines (SVMs) using a polynomial kernel, exactly as done in the original paper. To be completely clear, the training and testing part used to perform the experiments is the same for all the methods. The results, in terms of confusion matrix, are depicted in Fig. 6(a)-(b). It can be noted that our classifier achieves considerably better performances. In addition, we provide in Fig. 6(c) the confusion matrix for the enriched dataset. The performances are slightly below with respect to Fig. 6(b), because the enriched dataset is more challenging.



(a) Original, Orozco et al. (2009)

(b) Original, ARCO

(c) Enriched, ARCO

Figure 6: On the top row, some examples of the 4 Head original dataset by (Orozco et al., 2009). On the last row, the confusion matrixes for the head orientation classification: (a) (Orozco et al., 2009) and (b) ARCO comparisons on the original dataset, and (c) ARCO on the enriched dataset.

As to the analysis of social exchanges, we shot a video of about 3 hours and a half, portraying a vending machines area where students take coffee and discuss. The video footage was acquired with a monocular IP camera, located on a upper angle of the room. The people involved in the experiments were not aware of the aim of the experiments, and behaved naturally. Afterwards, since creating the ground truth by using only the video is an hard task, we asked to some of them to fill a questionnaire inquiring if they talked to someone in the room and to whom. Then, a video analysis was performed by a psychologist able to detect the presence of interactions between people. The questionnaires were used as supplementary material to confirm the validity of the generated ground truth. This offers us a more trustworthy set of ground truth data for our experiments.

We picked 12 subsequences of about 2 minutes each[3]. The 3.5h video has been reduced to this small set of sequences for several reasons: first, a lot of frames are empty, because the recording has been done on the early morning. Second, we have used only the sequences where the ground truth was evident and clear, *i.e.*, we know the components of each group. Third, they were chosen such that to represent different situations, with people talking in groups[4] and other people not interacting with anyone.

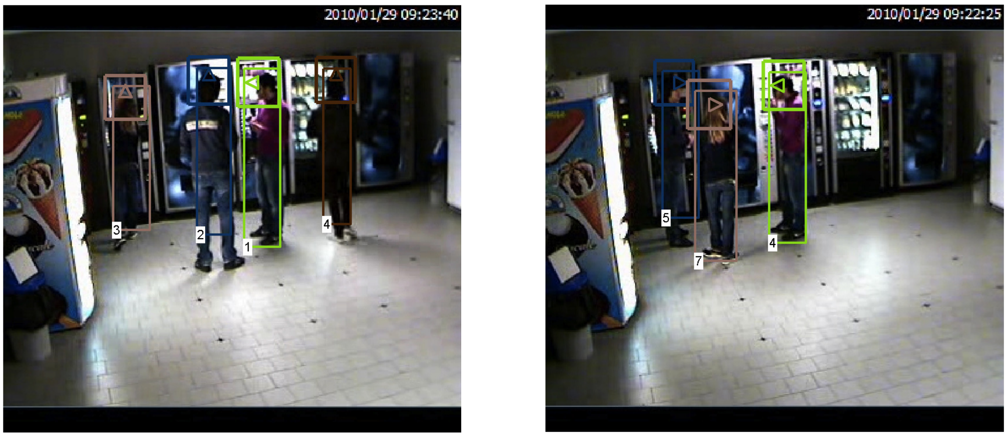


Figure 7: Examples of tracking and head orientation classification results. The biggest box represents the tracking estimation, the smaller box the area where the head is positioned, and the triangle depicts the estimated head orientation.

For each subsequence, we estimate tracking, head orientation classification (some examples are shown in Figure 7) and we build the three-dimensional IRPM, that tells which people are potentially interacting at a specific moment. Please, note that for the head classification part we enriched the 4 Head Pose dataset with head images coming from the Vending Machine dataset, in order to enrich accuracy and robustness. We added about 150 images for each FG class, and 1840 images to the Background.

[3] The dataset is downloadable from `http://www.lorisbazzani.info/code-datasets/multi-camera-dataset/`

[4] The groups are formed by 3 individuals, in average.

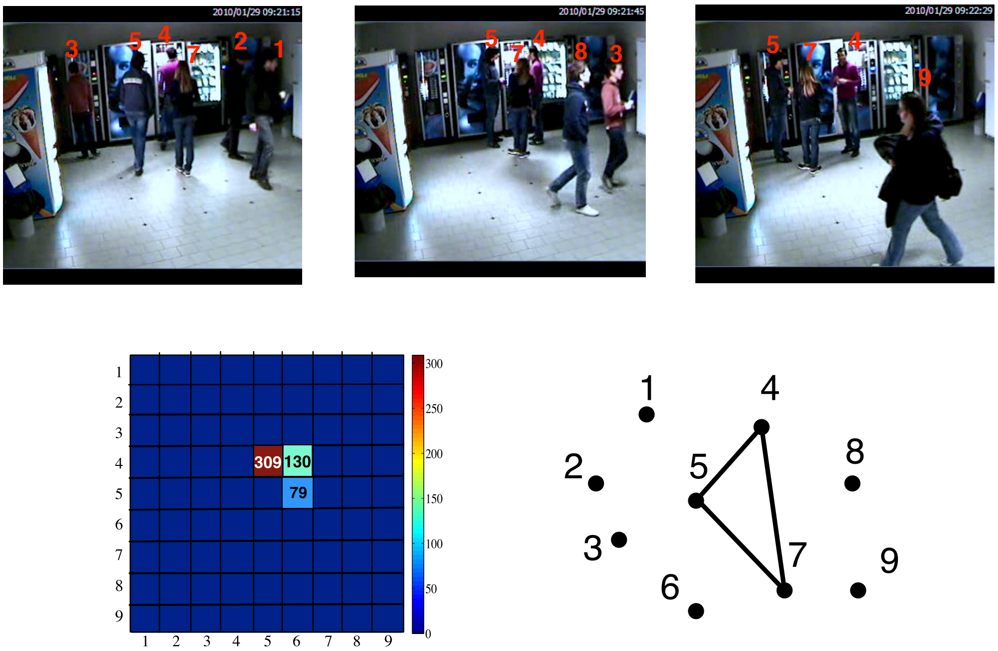


Figure 8: Example of condensed IRPM analysis of sequence S_{04}. On the top row, some frames of the sequence. On the bottom row, on the left, the thresholded cIRPM matrix. Being the cIRPMs symmetric and having null main diagonals, we report for clarity only its strictly upper triangular part. On the right, the correspondent graph. As you can notice, only one group (composed by people 4, 5 and 7) is detected. This is correct, since the other people of the sequence do not interact.

We compare our results with the ground truth. 8/12 sequences where correctly interpreted by our system. One can be considered wrong, because there are 2 groups in the scene, and our system reveals that they belong all to the same group. In the other three sequences there are some imprecisions, like a person left out of a group. These imprecisions are mainly due to error propagation from tracking and head orientation classification, particularly challenging when people are grouped together and frequently intersect. A qualitative analysis of the results is shown in Figures 8, 9 and 10. The first row of each figure depicts three sampled frames from each sequence and contains the identifiers of each person. The second row depicts the $cIRPM$ on the left[5] and the graph structure that defines the group interactions on the right. In these three experiments, all the groups are detected correctly; Fig. 10 shows that our model is able to detect interactions when the scene

[5] Blue cells mean zeros. The values of the $cIRPM$ below Th are discarded.

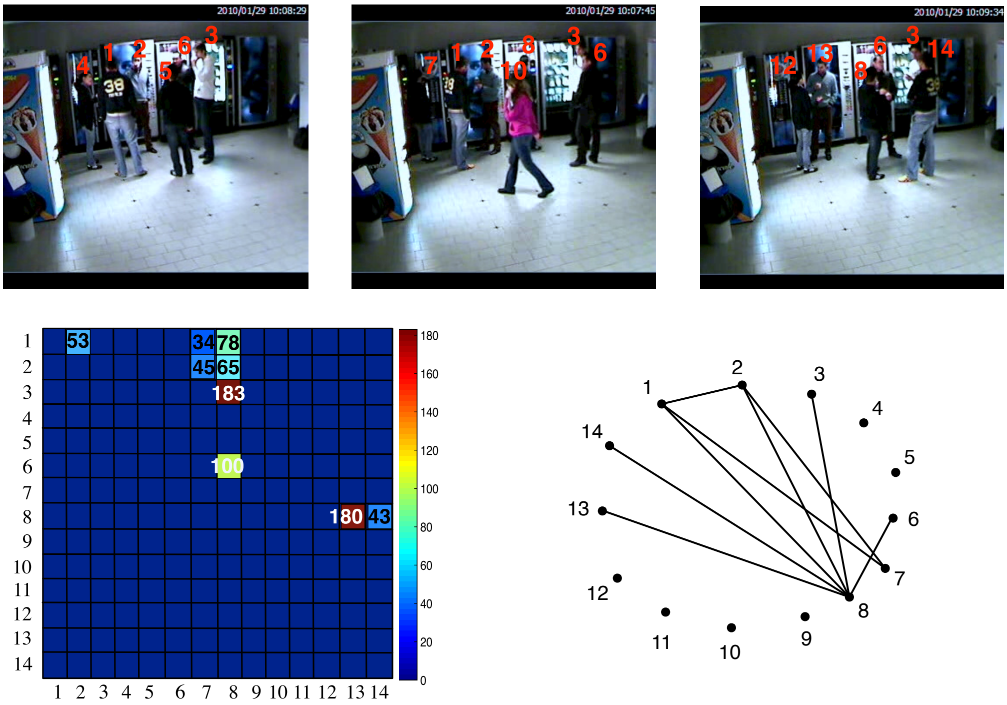


Figure 9: Example of condensed IRPM analysis of sequence S_{08}. On the top row, some frames of the sequence. On the bottom row, on the left, the thresholded cIRPM matrix. On the right, the correspondent graph. One big group (1,2,3,6,7,8,13,14) is detected. Please, note that some people are represented by more than one track, since due to severe or complete occlusions the tracks are sometimes lost and reinitialized. The group selected is correctly composed by the people associated to the labels. Another person (10) enters in the room and does not interact. The same behavior is witnessed in the cIRPM.

contains several groups.

A more sophisticated analysis of accuracy performances of our method is shown in Fig. 11 and Fig. 12. The graphs summarize the group detection accuracy in terms of precision (on the left) and recall (on the right). In the definition of those measurements, we consider as true positive when a group is detected considering all its constitutive members. If a person that belongs to a group is not detected, we have a false negative, and a similar reasoning applies for the false positive.

Fig. 11 depicts the statistics increasing the size T of the time interval $[t-T+1, T]$ (x-axis) used to accumulate the IRPM. Each curve corresponds to a value of threshold Th (5, 20, 60 and 100). From this figure, we notice that first of all increasing T gives worse accuracy. Moreover, the peak of

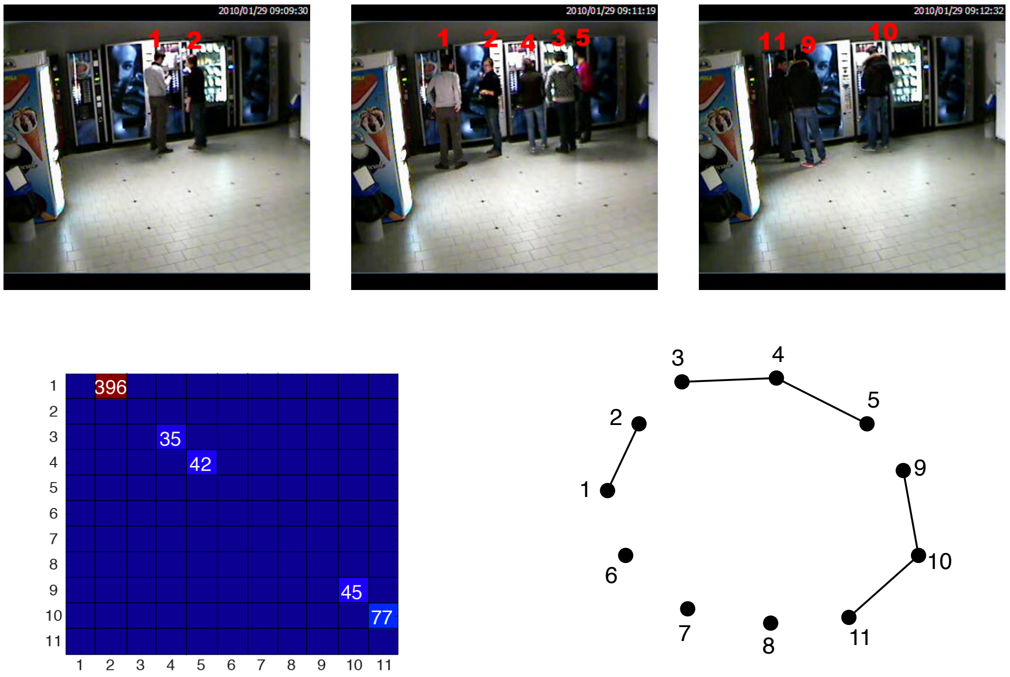


Figure 10: Example of condensed IRPM analysis of sequence S_{01}. On the top row, some frames of the sequence. On the bottom row, on the left, the thresholded cIRPM matrix. On the right, the correspondent graph. Three groups (1,2),(3,4,5), and (9,10,11) are detected. Please, note that some people are represented by more than one track, since due to severe or complete occlusions the tracks are sometimes lost and reinitialized (e.g. 6,7,8 are reinitialized as 9,10,11, respectively).

each curve depends on both the threshold and the time interval size. We obtain the best performances by setting the Th equal to 20; the peak of this curve corresponds to a T equal to 300. Instead, Fig. 12 shows the performances increasing the threshold (x-axis) used to detect the groups. Each curve corresponds to a value of T (120, 300, 480, 720, 900, and 1200). The common behavior of all the curves is that increasing and decreasing too much the threshold decreases the accuracy. This analysis confirms that the best performances are given by setting the threshold to 20 and the time interval to 300. When T increases the accuracy drastically decreases and the peak of each curve is shifted, depending by the time interval size.

Intuitively, when the threshold is too low and the time window is too small, our method detects interactions that could contain false positive. Increasing the size of the time window and the threshold permits to average out and cancel out these false positive, because the IRPM becomes more stable. On the other hand, when the threshold is too high, our model is not able to detect interactions, because $cIRPM_T(i,j) > Th$ is zero for each (i,j). To deal with this problem, we could make the time interval larger. However, in this case, a group interaction interval could be smaller than the time window, and in any case the threshold is too high to detect groups. For these reasons, precision and recall in Fig. 11 and Fig. 12 decrease before and after the optimal setting of the parameters ($Th = 20$ and $T = 300$).

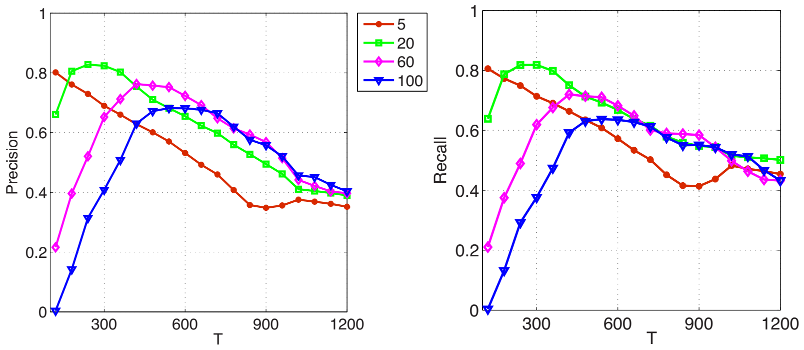


Figure 11: Evaluation of precision (left) and recall (right) of the proposed method varying the size of the time interval $[t - T + 1, t]$ (x-axis) used to compute the IRPM. The graph shows one curve for each threshold (5, 20, 60 and 100). The maximum for both the statistics is given by setting $Th = 20$.

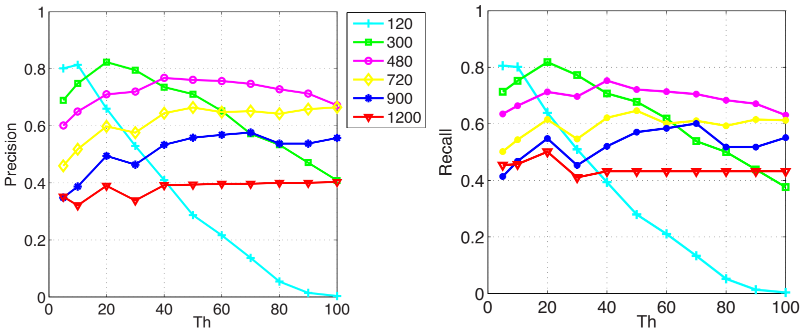


Figure 12: Evaluation of precision (left) and recall (right) of the proposed method varying the threshold Th (x-axis) used to detect the groups. The graph shows one curve for each time window (120, 300, 480, 720, 900, and 1200). The maximum for both the statistics is given by setting $T = 300$ and the peak is where $Th = 20$ (also accordingly with Fig. 11).

6. Conclusions

In this paper, we proposed a novel framework which may help in understanding 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 also possible to jointly investigate 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 funded by the FP7 EU-Project SAMURAI, grant FP7-SEC- 2007-01 No. 217899.

Authors' biosketches

- **Loris Bazzani** received the Laurea degree in Information Technology and Laurea Specialistica degree in Intelligent and Multitimedia Systems from University of Verona, Italy, in 2006 and 2008, respectively. He is currently a Ph.D. student in the Department of Computer Science, University of Verona working with the Vision, Image Processing and Sounds (VIPS) Lab. He was a visiting Ph.D. student in 2010 at the Laboratory of Computational Intelligence, University of British Columbia, Vancouver. His research interests include computer vision, machine learning, and sequential Monte Carlo methods with application to automatic video-surveillance. He is a student member IEEE.

- **Marco Cristani** received the Laurea degree in 2002 and the Ph.D. degree in 2006, both in computer science from the University of Verona, Verona, Italy. He was a visiting Ph.D. student at the Computer Vision Lab, Institute for Robotics and Intelligent Systems School of Engineering (IRIS), University of Southern California, Los Angeles, in 2004-2005. He is now Assistant Professor with the Dipartimento di Informatica, University of Verona, working with the Vision, Image Processing and Sounds (VIPS) Lab. He is also Team Leader with the

Istituto Italiano di Tecnologia, Genova, working with the PLUS Lab. His main research interests include statistical pattern recognition, generative modeling via graphical models, and nonparametric data fusion techniques, with applications on surveillance, segmentation, and image and video retrieval. He is the author of several papers in the above subjects and a reviewer for several international conferences and journals. He is also organizer of Workshops and PhD schools. He is IEEE and IAPR member.

- **Diego Tosato** received the Laurea degree in Information Technology and Laurea Specialistica degree in Intelligent and Multimedia Systems from the University of Verona, Italy, in 2006 and 2008. He is currently a Ph.D. student in the Department of Computer Science, University of Verona working with the Vision, Image Processing and Sounds (VIPS) Lab. His research interests are in computer vision, machine learning, and statistical methods to image understanding problems.

- **Michela Farenzena** was born in Verona - Italy - in 1978. She received her Laurea(Master) degree in Computer Science from the University of Verona in 2003. She received the Dottorato di Ricerca (PhD) in Computer Science from the University of Verona in 2007. As a PhD student she had been Visiting Doctoral student at Queen Mary University, London in 2005. She had been Postgraduate fellow at LASMEA, Université Blaise Pascal, Clermont-Ferrand - France - in collaboration with the Commissariat de l'Energie Atomique (CEA) in 2008. She is now working as Computer Vision specialist in eVS - embedded Vision Systems. Her research is focused on various topics in Computer Vision and Image Analysis.

- **Giulia Paggetti** received the Laurea degree in General and Experimental Psychology in 2005 and the Master degree in Experimental Psychology in 2008 from the University of Firenze. She is currently a Ph.D. student in the Department of Computer Science, University of Verona working with the Vision, Image Processing and Sounds (VIPS) Lab. Her research interests concern perceptual imaging as well as analysis of nonverbal-social behavior in real world.

- **Gloria Menegaz** obtained the MSc. degree in Electronic Engineering at the Polytechnic University of Milan (Italy) in 1993, the MSc.

in Information Technology at the Center for Research and Education in Information Technology (CEFRIEL) Polytechnic University of Milan in 1994 and the Ph.D. in Applied Science at the Signal Processing Institute (ITS) of the Swiss Federal Institute of Technology (EPFL) in July 2000. From May to October 2000 she was with the Biomedical Imaging Group (BIG), headed by Prof. Michael Unser, as post-doctoral fellow. From 10/2000 to 10/2002 she was First Research Assistant at the Audiovisual Communications Lab. (LCAV) of EPFL, leaded by Prof. Martin Vetterli. From 12/2002 to 03/2004 she was Assistant Professor (Maitre Assistante) at the Department of Computer Science of the University of Fribourg (Switzerland). In March 2004 she has been awarded of a grant Brain Drain professorship from the Italian Ministry of University and Research (MIUR) for joining Department of Information Engineering of the University of Siena (Italy) as an adjunct Professor. Since Oct. 2007 she is Associate Professor at the Department of Computing of the Faculty of Sciences of the University of Verona (Italy). Her research interests are in the field of perceptual image processing, computational vision, visual perception and model-based processing of multidimensional data.

- **Vittorio Murino** received the Laurea degree in electronic engineering in 1989 and the Ph.D. degree in electronic engineering and computer science in 1993, both from the University of Genoa, Genoa, Italy. He is a Full Professor and Chairman of the Department of Computer Science, University of Verona. From 1993 to 1995, he was a Postdoctoral Fellow in the Signal Processing and Understanding Group, Department of Biophysical and Electronic Engineering, University of Genova, where he supervised of research activities on image processing for object recognition and pattern classification in underwater environments. From 1995 to 1998, he was an Assistant Professor of the Department of Mathematics and Computer Science, University of Udine, Udine, Italy. Since 1998, he has been with the University of Verona, where he founded and is responsible for the Vision, Image Processing, and Sound (VIPS) Laboratory. He is also the founder and director of the Pattern analysis, Learning, and image Understanding laboratory (Plus) Lab with the Istituto Italiano di Tecnologia, Genova. He is scientifically responsible for several national and European projects and is an Evaluator for the European Commission of research project proposals related to different scientific programmes and frameworks.

His main research interests include computer vision and pattern recognition, probabilistic techniques for image and video processing, and methods for integrating graphics and vision. He is author or co-author of more than 150 papers published in refereed journals and international conferences. Dr. Murino is a referee for several international journals, a member of the technical committees for several conferences (ECCV, ICPR, ICIP), and a member of the editorial board of Pattern Recognition, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, Pattern Analysis and Applications, and Electronic Letters on Computer Vision and Image Analysis (ELCVIA). He was the promotor and Guest Editor of four special issues of Pattern Recognition and is a Fellow of the IAPR.

References

Ambady, N., Rosenthal, R., 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta analysis. Psychological bulletin 111 (2), 256–274.

Ba, S. O., marc Odobez, J., 2006. A study on visual focus of attention recognition from head pose in a meeting room. In: Machine Learning for Multimodal Interaction. pp. 75–87.

Benezeth, Y., Jodoin, P., Emile, B., Laurent, H., Rosenberger, C., 2008. Review and evaluation of commonly-implemented background subtraction algorithms. In: Proceedings of International Conference on Pattern Recognition. pp. 1–4.

Benfold, B., Reid, I., September 2009. Guiding visual surveillance by tracking human attention. In: Proceedings of the 20th British Machine Vision Conference.

Blackman, S. S., 2004. Multiple hypothesis tracking for multiple target tracking. IEEE Aerospace and Electronic Systems Magazine 19 (1), 5–18.

Breiman, L., Friedman, J., Olshen, R., Stone, C., Breiman, L., Hoeffding, W., Serfling, R., Friedman, J., Hall, O., Buhlmann, P., et al., 1984. Classification and Regression Trees. Ann. Math. Statist. 19, 293–325.

Chellappa, R., Roy-Chowdhury, A., Zhou, S., 2005. Recognition of humans and their activities using video. Synthesis Lectures on Image, Video & Multimedia Processing 1 (1), 1–173.

Chen, Y., Chen, C., Hung, Y., Chang, K., 2009. Multi-class multi-instance boosting for part-based human detection. In: IEEE 12th International Conference on Computer Vision Workshops. pp. 1177–1184.

Choudhury, T., Pentland, A., Nov. 2002. The sociometer: A wearable device for understanding human networks. In: Computer Supported Cooperative Work - Workshop on Ad hoc Communications and Collaboration in Ubiquitous Computing Environments. pp. 1–6.

Cristani, M., V.Murino, Vinciarelli, A., 2010. Socially intelligent surveillance and monitoring: Analysing social dimensions of physical space. In: First IEEE International Workshop on Socially Intelligent Surveillanceand Monitoring. San Francisco, California, pp. 51–58.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Vol. 1. pp. 886–893.

Farenzena, M., Bazzani, L., Murino, V., Cristani, M., 2009a. Towards a subject-centered analysis for automated video surveillance. In: Proceedings of the 15th International Conference on Image Analysis and Processing. ICIAP '09. Springer-Verlag, Berlin, Heidelberg, pp. 481–489.

Farenzena, M., Fusiello, A., Gherardi, R., Toldo, R., 2008. Towards unsupervised reconstruction of architectural models. In: Proceedings of Vision, Modeling, and Visualization. pp. 41–50.

Farenzena, M., Tavano, A., Bazzani, L., Tosato, D., Paggetti, G., Menegaz, G., Murino, V., Cristani, M., 2009b. Social interactions by visual focus of attention in a three-dimensional environment. In: Workshop on Pattern Recognition and Artificial Intelligence for Human Behaviour Analysis. pp. 1–8.

Francois, A., Nevatia, R., Hobbs, J., Bolles, R., 2005. Verl: An ontology framework for representing and annotating video events. IEEE MultiMedia 12, 76–86.

Freeman, L., 1989. Social networks and the structure experiment. In: Research Methods in Social Network Analysis. pp. 11–40.

Freund, Y., Schapire, R., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55 (1), 119–139.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: A statistical view of boosting. The annals of statistics 28 (2), 337–374.

Fuentes, L., Velastin, S., 2006. People tracking in surveillance applications. Image and Vision Computing 24 (11), 1165 – 1171.

Huang, C., Ai, H., Li, Y., Lao, S., 2005. Vector boosting for rotation invariant multi-view face detection. In: Proceedings of the International Conference on Computer Vision. pp. 446–453.

Hung, H., Jayagopi, D. B., Ba, S., Odobez, J.-M., Gatica-Perez, D., 2008. Investigating automatic dominance estimation in groups from visual attention and speaking activity. In: Proceedings of the 10th international conference on Multimodal interfaces. ICMI '08. ACM, New York, NY, USA, pp. 233–236.

Jabarin, B., Wu, J., Vertegaal, R., Grigorov, L., 2003. Establishing remote conversations through eye contact with physical awareness proxies. In: CHI '03 extended abstracts on Human factors in computing systems. CHI EA '03. ACM, New York, NY, USA, pp. 948–949.

Lablack, A., Djeraba, C., 2008. Analysis of human behaviour in front of a target scene. In: Proceedings of the 19th International Conference on Pattern Recognition, 2008. IEEE, pp. 1–4.

Langton, S., Watt, R., Bruce, V., 2000. Do the eyes have it? cues to the direction of social attention. Trends in Cognitive Neuroscience 4 (2), 50–58.

Lanz, O., September 2006. Approximate bayesian multibody tracking. IEEE Trans. Pattern Anal. Mach. Intell. 28, 1436–1449.

Lanz, O., Brunelli, R., Chippendale, P., Voit, M., Stiefelhagen, R., 2009. Extracting interaction cues: Focus of attention, body pose, and gestures. Computers in the Human Interaction Loop, 87–93.

Li, S. Z., Zhang, Z., September 2004. Floatboost learning and statistical face detection.

IEEE Trans. Pattern Anal. Mach. Intell. 26, 1112–1123.

Li, S. Z., Zhu, L., Zhang, Z., Blake, A., Zhang, H., Shum, H., 2002. Statistical learning of multi-view face detection. In: Proceedings of the 7th European Conference on Computer Vision. Springer-Verlag, London, UK, pp. 67–81.

Liu, X., Krahnstoever, N., Ting, Y., Tu, P., 2007. What are customers looking at? In: Advanced Video and Signal Based Surveillance. pp. 405–410.

Marques, J. S., Jorge, P. M., Abrantes, A. J., Lemos, J. M., Jun 2003. Tracking groups of pedestrians in video sequences. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop. Vol. 9. pp. 101–107.

Matsumoto, Y., Ogasawara, T., Zelinsky, A., 2002. Behavior recognition based on head-pose and gaze direction measurement. In: Proc. Int'l Conf. Intelligent Robots and Systems. Vol. 4. pp. 2127–2132.

Mckenna, S. J., Jabri, S., Duric, Z., Wechsler, H., Rosenfeld, A., 2000. Tracking groups of people. Computer Vision and Image Understanding, 42–56.

Murphy-Chutorian, E., Trivedi, M. M., April 2009. Head pose estimation in computer vision: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 31, 607–626.

Orozco, J., Gong, S., Xiang, T., 2009. Head pose classification in Crowded Scenes. In: Proceedings of the British Machine Vision Conference.

Otsuka, K., Yamato, J., Takemae, Y., Murase, H., 2006. Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns. In: Proceedings of the Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 1175–1180.

Paisitkriangkrai, S., Shen, C., Zhang, J., 2008. Performance evaluation of local features in human classification and detection. Computer Vision, Institution of Engineering and Technology 2 (4), 236–246.

Panero, J., Zelnik, M., 1979. Human dimension & interior space: a source book of design reference standards. Whitney Library of Design.

Pantic, M., Pentland, A., Nijholt, A., 2009. Special issue on human computing. IEEE Trans. on on Systems, Man, and Cybernetics, Part B 39 (1), 3–6.

Pentland, A., January 2000. Looking at people: Sensing for ubiquitous and wearable computing. IEEE Trans. Pattern Anal. Mach. Intell. 22, 107–119.

Pentland, A., July 2007. Social signal processing. Signal Processing Magazine, IEEE 24 (4), 108–111.

Picard, R., 2000. Affective computing. The MIT press.

Preparata, F. P., Shamos, M. I., 1985. Computational geometry: an introduction. Springer-Verlag New York, Inc., New York, NY, USA.

Robertson, N., Reid, I., 2006. Estimating gaze direction from Low-Resolution faces in video. In: Proceedings of the IEEE European Conference on Computer Vision. pp. 402–415.

Schapire, R., Singer, Y., 1999. Improved boosting algorithms using confidence-rated predictions. Mach. Learn. 37, 297–336.

Smith, K., Ba, S., Odobez, J., Gatica-Perez, D., 2008. Tracking the visual focus of attention for a varying number of wandering people. IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (7), 1–18.

Smith, K., Gatica-Perez, D., Odobez, J., Ba, S., 2005. Evaluating multi-object tracking. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. pp. 36–43.

Smith, P., Shah, M., da Vitoria Lobo, N., 2003. Determining driver visual attention with one camera. IEEE Transactions on Intelligent Transportation Systems 4 (4), 205–218.

Snavely, N., Seitz, S., Szeliski, R., 2006. Photo tourism: exploring photo collections in 3d. In: ACM Transactions on Graphics. Vol. 25. ACM, pp. 835–846.

Stiefelhagen, R., Bowers, R., Fiscus, J. (Eds.), 2008. Multimodal Technologies for Perception of Humans: International Evaluation Workshops on Classification of Events, Activities and Relationships 2007. Springer-Verlag, Berlin, Heidelberg.

Stiefelhagen, R., Finke, M., Yang, J., Waibel, A., 1999. From gaze to focus of attention. In: Visual Information and Information Systems. pp. 761–768.

Stiefelhagen, R., Garofolo, J. (Eds.), 2007. Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities and Relationships 2006. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Stiefelhagen, R., Yang, J., Waibel, A., 2002. Modeling focus of attention for meeting indexing based on multiple cues. IEEE Transactions on Neural Networks 13, 928–938.

Tuzel, O., Porikli, F., Meer, P., 2006. Region covariance: A fast descriptor for detection and classification. Proceedings of the European Conference on Computer Vision, 589–600.

Tuzel, O., Porikli, F., Meer, P., October 2008. Pedestrian detection via classification on riemannian manifolds. IEEE Trans. Pattern Anal. Mach. Intell. 30, 1713–1727.

Vinciarelli, A., Pantic, M., Bourlard, H., 2009. Social Signal Processing: Survey of an emerging domain. Image and Vision Computing Journal 27 (12), 1743–1759.

Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 1. pp. 511–518.

Voit, M., Stiefelhagen, R., 2008. Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In: Proceedings of the 10th international conference on Multimodal interfaces. ICMI '08. ACM, New York, NY, USA, pp. 173–180.

Waibel, A., Schultz, T., Bett, M., Denecke, M., Malkin, R., Rogina, I., Stiefelhagen, R., 2003. SMaRT: the Smart Meeting Room task at ISL. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. pp. 752–755.

Whittaker, S., Frohlich, D., Daly-Jones, O., 1994. Informal workplace communication: what is it like and how might we support it? In: Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence. CHI '94. ACM, New York, NY, USA, pp. 131–137.

Wu, B., Ai, H., Huang, C., Lao, S., 2004. Fast rotation invariant multi-view face detection based on real adaboost. In: Proceedings of the Sixth IEEE international conference on Automatic face and gesture recognition. FGR' 04. IEEE Computer Society, Washington, DC, USA, pp. 79–84.

Wu, B., Nevatia, R., 2008. Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection. In: Proceedings of the International Conference of Computer Vision and Pattern Recognition.

Wu, B., Nevatia, R., April 2009. Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses. International

Journal of Computer Vision 82, 185–204.

Zheng, W., Gong, S., Xiang, T., 2009. Associating groups of people. In: Proceedings of the British Machine Vision Conference.