Towards a Subject-Centered Analysis for Automated Video Surveillance

Michela Farenzena[1], Loris Bazzani[1],
Vittorio Murino[2], and Marco Cristani[2]

[1] Dipartimento di Informatica, Università di Verona, Italy
[2] IIT, Istituto Italiano di Tecnologia, Genova, Italy

Abstract. In a typical video surveillance framework, a single camera or a set of cameras monitor a scene in which human activities are carried out. In this paper, we propose a complementary framework where human activities can be analyzed under a subjective point of view. The idea is to represent the focus of attention of a person in the form of a 3D view frustum, and to insert it in a 3D representation of the scene. This leads to novel inferences and reasoning on the scene and the people acting in it. As a particular application of this proposed framework, we collect the information from the subjective view frusta in an Interest Map, i.e. a map that gathers in an effective and intuitive way which parts of the scene are observed more often in a defined time interval. The experimental results on standard benchmark data witness the goodness of the proposed framework, encouraging further efforts for the development of novel applications in the same direction.

1 Introduction

The visual focus of attention (VFOA) is a well-studied phenomenon in psychological literature, recently employed as strong social signal through which it is possible to robustly reckon social interest or the presence of an ongoing dialog. Social signaling aims to embed these psychological findings with computer science methods, in order to give robust insight about social activities (see [1] for a review). In this field, the use of the gaze direction as computational feature for the VFOA is widely accepted and tested. However, the main approaches in this area [2,3] focus mostly on meeting scenarios, in order to discover the addressee of a particular person, i.e. the subject one wants to address for a conversational exchange. Recently, Smith et al. [4] extend the VFOA problem to a wider scenario, estimating the gaze direction of people wondering freely in an outdoor scene. This paper goes in the same direction, widening the use of VFOA to the contest of automated video surveillance.

The main contribution of this paper is to translate the notion of VFOA into a feature that can be employed in a surveillance scenario. We claim that the three-dimensional (3D) space, not employed in [4], is the appropriate environment where to reason about this issue. Assuming that the monitoring camera is calibrated, we propose a novel feature, called *Subjective View Frustum* (SVF).

P. Foggia, C. Sansone, and M. Vento (Eds.): ICIAP 2009, LNCS 5716, pp. 481–489, 2009.

The SVF is an estimation of the view frustum of a single person. It is modeled as a 3D polyhedral whose volume represents the portion of the scene in which the subject is reasonably focused on. Having a rough 3D map of the scene being monitored, we consistently locate each person and its related SVF in the 3D space. Hence, we are able to satisfactorily analyze people attention in wide contexts, where several, overlapping people are present and in which the position of the video sensor does not permit a fully detailed acquisition of the human head.

More generally, the proposed framework provides a point of analysis that is different and complementary to the standard, third person point of view of a single camera or a set of cameras. This opens new perspectives on several interesting applications and inferences, in the direction of a subjects-centered understanding of human activities in surveillance contexts.

The second contribution of the paper is a visualization application of the proposed SFV-based framework, called the *Interest Map*. Since the part of the scene that intersects the SVF is the scene observed by the SVF's owner, we collect this information for each subject, over a given time interval. This permits to infer which are the parts of the scene that are more observed, thus more plausibly subjected to attention. The gathered information is visualized as a suitably colored map, in which hot colors represent the areas more frequently observed, vice versa for "cold" areas. This kind of inference is highly informative, at least for two reasons. The first one is diagnostics, in the sense that it gives us the possibility to observe which are the areas of a scene that arouse more attention in the people. The other one is prognostics, since it enables us to devise the parts of the scene that are naturally more seen, because for example they are the natural front of view in a narrow transit area, or for other reasons that this method cannot guess (the Interest Map only highlights the tangible effects). This application could be employed for a posteriori analysis. In a museum, for example, one may be interested in understanding which artworks receive more attention, or in a market which areas attract more the customers. In a prognostic sense it may be useful for marketing purposes, such as for example decide where to hang an advertisement.

A basic element of our scenario is a tracking module that is able to follow multiple objects, and deal with occlusions and overlapping people. It is based on the multi-object particle filtering platform proposed in [5] and it is described in Section 2. After that, Section 3 introduces the SVF model and explains how it is estimated. Section 4 describes how to build the Interest Map. Then we show the experimental results in Section 5 and we draw our conclusions in Section 6.

2 Hybrid Joint Separable Particle Filtering Tracker

The tracking module we employ is based on a particle filtering strategy. Particle filters offer a probabilistic framework for recursive dynamic state estimation. The approach was born originally for single-object tracking [6], then later it was extended to a multi-object tracking scenario [7]. Multi-object particle filters follow different strategies to achieve good tracking performances avoiding huge computational burdens. These are due primarily to the high number of particles required, which is (in general) exponential in the number of objects to track. Recently, an interesting yet general solution has been proposed by Lanz in [5]. He defined the Hybrid Joint-Separable (HJS) filter, that maintains a linear relationship between number of objects and particles.

The goal is to determine the posterior distribution $p(x_t|z_{1:t})$, where x_t is the current state, z_t is the current measurement, and $x_{1:t}$ and $z_{1:t}$ are respectively the states and the measurements up to time t. We refer to x_t as the state of single object, and $\mathbf{x}_t = \{x_t^1, x_t^2, \ldots, x_t^K\}$ the joint state (all objects). Finally, the posterior distribution $p(x_t|z_{1:t})$ is approximated by a set of N weighted particles, *i.e.* $\{(x_t^n, w_t^n)\}_{n=1}^N$.

The HJS approach represents a theoretical grounded compromise between dealing with a strictly joint process and instantiating a single, independent tracking filter for each distinct object. In practice, HJS alternates a separate modeling during the sampling step and a hybrid joint-separate formulation in the dynamical and observational steps.

The rule that permits the crossing over joint-separable treatments is based on the following approximation (see [5] for rigorous math details):

$$p(\mathbf{x}_t|z_{1:\tau}) \approx \prod_k p(x_t^k|z_{1:\tau}) \tag{1}$$

that is, the joint posterior could be approximated by the product of its marginal components. This assumption allows to sample the particles in the single state space (thus requiring a linear proportionality between number of object and number of samples), and to update the weights in the joint state space. The updating exploits a) a joint dynamical model that builds the distribution $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, explaining how the system does evolve, and b) a joint observational model that provides estimates for the distribution $p(z_t|\mathbf{x}_t)$, explaining how the observations can be related to the state of the system. Both the models take into account the interactions among objects. In particular $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ accounts for physical interactions between targets, thus avoiding track coalescence of spatially near targets.

The observational model $p(z_t|\mathbf{x}_t)$ quantifies the likelihood of the single measure z_t given the state $\mathbf{x}_t$, considering inter-objects occlusions. It is built upon the representation of the targets, that here are constrained to be human beings. The human body is represented by its three components: head, torso and legs. The observational model works by evaluating a separate appearance score for each object (summing then the contribute of the single parts). This score is encoded by a distance between the histograms of the model and the hypothesis (a sample), and it involves also a joint reasoning captured by an *occlusion map*. The occlusion map is a 2D projection of the 3D scene which focuses on the particular object under analysis, giving insight on what are the expected visible portions of that object. This is obtained by exploiting the hybrid particles set $\{x_p\}_{p=1}^{NK}$ in an incremental visit procedure on the ground floor. The hypothesis nearest to the camera is evaluated first. Its presence determines an occluding cone in the scene, with an associated confidence that depends on the observational likelihood achieved. Parts of other objects deeper in the scene that fall in the occlusion cone are considered less in their observational likelihood computation. The process of map building is iterated by going deeper in the scene.

In formulae, the observation model is defined as

$$p(z_t|x_p) \propto \exp\left(-\frac{\mathrm{fc}_p + \mathrm{bc}_p}{2\,\sigma^2}\right), \tag{2}$$

where fc_p is the foreground term, *i.e.*, the likelihood that an object matches to the model considering the unoccluded parts, and bc_p, the background term, accounts for the occluded parts of an object.

3 3D Modeling of the Subjective View Frustum

Once the tracks for one frame are available, the viewing direction of each tracked subject is derived and the information about the parts of the scene watched by him/her are suitably inserted in an accumulation matrix. This matrix will represent our Interest Map at the end of the computation. This reasoning goes by a sequence of 3D operations and the definition of the *Subjective View Frustum*, as detailed in the following.

3.1 3D Map Estimation

In this paper we suppose that the camera monitoring the area is fully calibrated, i.e. both internal parameters and camera position and orientation are known. For convenience, the world reference system is put on the ground floor, 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 floor is known. In fact, if P is the camera projection matrix and $\mathbf{M} = (M_x, M_y, M_z)$ the coordinates of a 3D point, the projection of $\mathbf{M}$ through P is given by two equations:

$$u = \frac{\mathbf{p}_1^\top \mathbf{M}}{\mathbf{p}_3^\top \mathbf{M}}, \quad v = \frac{\mathbf{p}_2^\top \mathbf{M}}{\mathbf{p}_3^\top \mathbf{M}}, \quad \text{with } \mathsf{P} = \begin{bmatrix} \mathbf{p}_1^\top \\ \mathbf{p}_2^\top \\ \mathbf{p}_3^\top \end{bmatrix}. \tag{3}$$

(u, v) are the coordinates of the image point. Thus, knowing (u, v) and M_z it is possible to estimate the position of $\mathbf{M}$ in the 3D space.

A rough reconstruction of the area, made up of the principal planes present in the scene, can therefore be carried out. An example in shown in Figure 1. These planes represent the areas of the scene that are interesting to analyze. The Interest Map will be estimated on them only. 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 (SfM) algorithm [8,9] is applied. In any case, this operation must be executed just once.

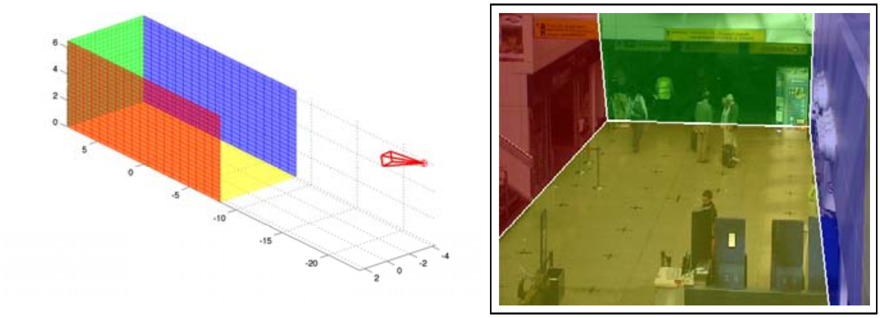

Fig. 1. 3D reconstruction of the area being monitored. On the left, the 3D map of the principal planes. The red cone represents the camera. On the right, the planes are projected through the camera and superimposed on one image.

3.2 Head Orientation Estimation

The tracking algorithm provides the position (x_{it}, y_{it}) of each person i present in the scene at a certain moment t. We need to calculate the head orientation in order to decide in which direction a person is looking. At the scale of a typical video surveillance scenario, tracking head direction is very difficult. Thus, are content with a rough estimator, that distinguishes among four possible directions (North, South, East, West) relative to the camera orientation.

In this paper we exploit the tracking information, i.e. the position of each person over the time. We calculate the angle between the motion direction and the camera orientation. The underlying assumption here is that the subject is looking in the direction towards which he/she is moving. This assumption is quite strong. It could be relaxed by employing a more sophisticated algorithm for head detection, such as [10], but it is beyond the purposes of this paper.

3.3 Subjective View Frustum

Once the view direction has been detected, a view frustum can be estimated. It represents the portion of 3D space seen by the subject. We call this portion *Subjective View Frustum* (SVF). Geometrically, we model the SVF as the polyhedron $\mathcal{D}$ depicted in Figure 2. 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 60° horizontally and 120° vertically. If we take into account the maximum field of view of a human, the SVF should be much bigger (around 140° on both directions). However, we considered that the focus of attention, especially when a person is moving, reduces the actual field of view.

The 3D coordinates of the points corresponding to the head and the feet of a subject are obtained from the (x_{it}, y_{it}) coordinates given by the tracker, under the assumption that he/she walks on the ground floor and is 1.8 m tall.

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

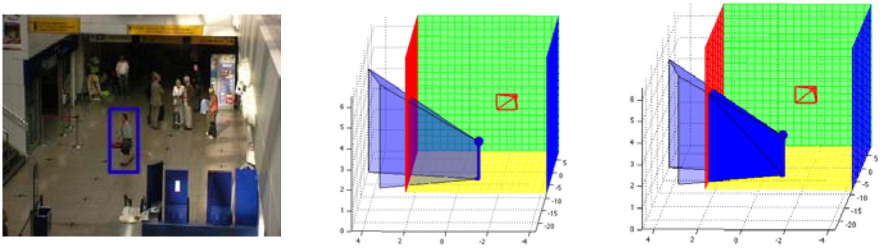

Fig. 2. SVF of the person detected in the frame on the left (blue square). At the center, a perspective view of the SVF (semi-transparent blue); on the right, the actual SVF inside the scene (solid blue).

their supporting planes respectively of the left, right and top side of the subject. Moreover, the SVF is also limited by the planes that set up the scene, according to the 3D map. The scene volume is similarly modeled as intersection of negative half-space. Thus, the exact SVF inside the scene can be computed solving a simple *vertex enumeration* problem, for which very efficient algorithms exists in literature [11].

4 Creation of the Interest Map

The SVF $\mathcal{D}$ represents the portion of 3D space seen by the subject. As mentioned before, we decided to concentrate our attention on the scene main planes only. A full volumetric reasoning could be tackled too, but this would capture other kinds of information, such as people interactions.

In order to record the SVF information, we project the SVF volume on each scene plane. This is equivalent to estimate the vertices of $\mathcal{D}$ lying on each plane, project these vertices on the image and select those pixels that lay inside the convex hull of the projected vertices. In this way the selected pixels can be inserted in an accumulation matrix, that is a 2D matrix of the same size of the camera frames. This also implies that the accumulation matrix is registered to the camera viewpoint. Two examples of this projection operation are shown in

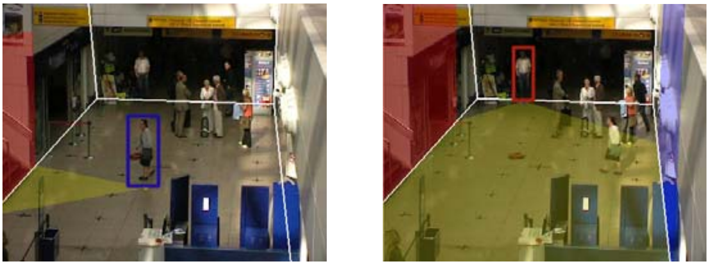

Fig. 3. Two examples of projection of the SVF on the scene main planes. The 3D map permits to suitably model the interactions of the SVF with the scene.

Figure 3. The contributions provided by all tracked people in the whole sequence, or a set of sequences, are conveyed in the same accumulation matrix. This matrix, at the end of the observation time window, is our Interest Map.

5 Experimental Results

We perform some tests over the PETS 2007 sequence sets. This aims to show the expressiveness of our framework on widely known and used datasets. 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 is captured by Camera 2, the second one is captured by Camera 4.

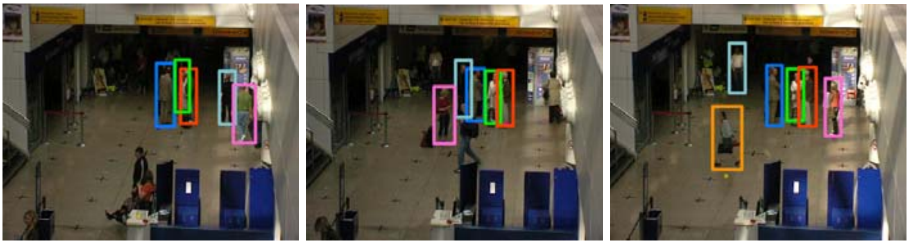

Fig. 4. Some frames of the sequence from camera 2. The bounding boxes highlight the tracking results.

In Figure 4 we show some frames of the first sequence, with highlighted the tracking results. Totally, 1 minute of activity has been monitored, tracking continuously averagely 5 people at a time.

The resulting Interest Map is depicted in Figure 5, superimposed as transparency mask to an image of the scene. From this map interesting considerations can be assessed. The "hottest" area in the one closest to the camera, in the direction of the stairs on the left. Indeed in the sequence many people cross that area from right to left. Another interesting area is at the end of the corridor, while the entrance on the left end has never been watched. Indeed the other people detected throughout the sequence are on the right end, going north.

For the second sequence, captured by Camera 4, one minute is monitored, tracking averagely 4 people at a time. The SVF analysis produces the results shown in Figure 6. In this case the most seen area is the left end corner of the corridor. Indeed most of the people in the sequence give the back to the camera. The other "hot" area is the left front corner, due to a person loitering there most of the time interval considered. As a comparison we plot together (right picture of Figure 6) the tracking results. This representation is less meaningful from the point of view of people attention analysis. Our information visualization technique is instead intuitive and it captures in a very simple and richer way where people attention is focused.

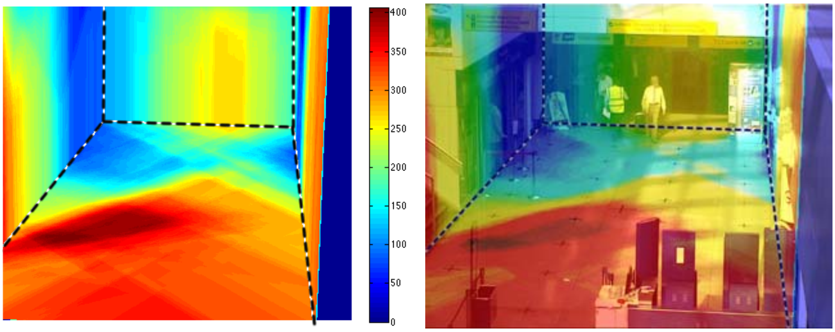

Fig. 5. On the left, the Interest Map for S07 sequence from camera 2. On the right, the same Interest Map superimposed on one frame of the sequence.

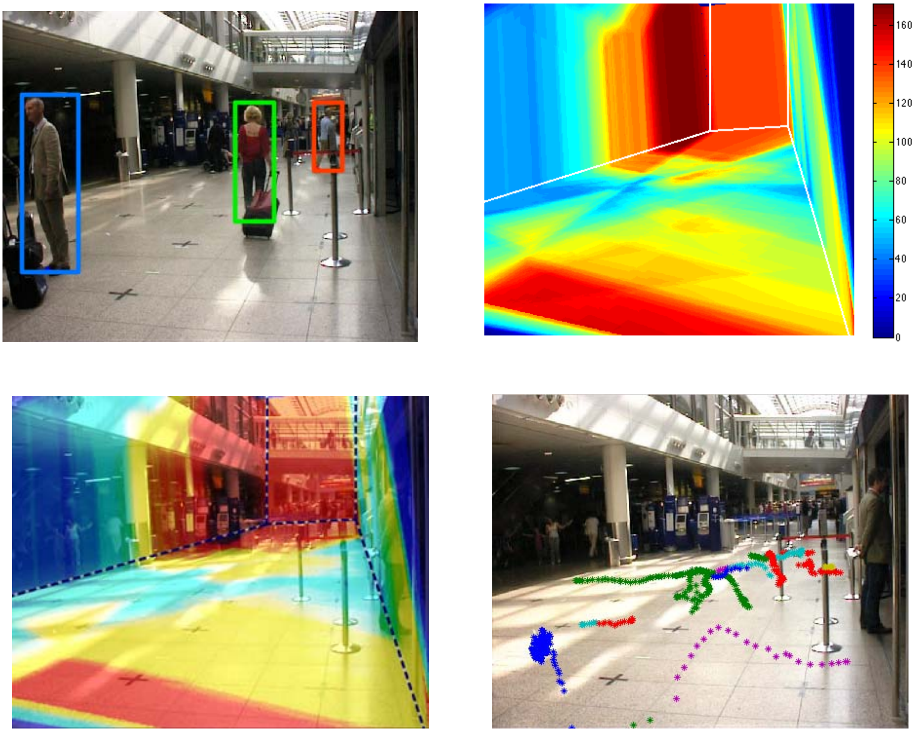

Fig. 6. On the left top, one frame of the second sequence, with highlighted the tracked people. On the right top, the Interest Map obtained. On the left bottom, the same Interest Map superimposed on one frame. On the right bottom, all the tracks estimated throughout the sequence displayed in the same frame.

6 Conclusions and Future Work

In this paper we proposed a complementary video surveillance framework focused on the subjective focus of attention. We showed that the 3D space is the appropriate environment for this issue: we model the view frustum of each person moving in the scene as a 3D polyhedral whose volume represents the portion of the scene in which the subject is reasonably focused on. As a particular application of this framework, we collect the focus of attention information in an Interest Map that gathers in an effective and intuitive way the information about the parts of the scene observed more in a defined time interval.

An interesting development toward a finer estimation of Interest Maps will be to employ a robust head pose estimator. Moreover, we plan to apply the SVF feature to investigate people interactions in wide areas.

Acknowledgements

This research is founded by the EU-Project FP7 SAMURAI, grant FP7-SEC-2007-01 No. 217899.

References

1. Gatica-Perez, D.: Automatic nonverbal analysis of social interaction in small groups: a review. Image and Vision Computing, Special Issue on Human Naturalistic Behavior (accepted for publication)
2. Jayagopi, D., Hung, H., Yeo, C., Gatica-Perez, D.: Modeling dominance in group conversations from nonverbal activity cues. IEEE Trans. on Audio, Speech, and Language Processing, Special Issue on Multimodal Processing for Speech-based Interactions 3(3) (2009)
3. Paul, C., Oswald, L.: Optimised meeting recording and annotation using real-time video analysis. In: Popescu-Belis, A., Stiefelhagen, R. (eds.) MLMI 2008. LNCS, vol. 5237, pp. 50–61. Springer, Heidelberg (2008)
4. Smith, K., Ba, S.O., Odobez, J.M., Gatica-Perez, D.: 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 (2008)
5. Lanz, O.: Approximate bayesian multibody tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(9), 1436–1449 (2006)
6. Isard, M., Blake, A.: Condensation: Conditional density propagation for visual tracking. Int. J. of Computer Vision 29, 5–28 (1998)
7. Isard, M., MacCormick, J.: Bramble: A bayesian multiple-blob tracker (2001)
8. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3D. In: SIGGRAPH Conference Proceedings, NY, USA, pp. 835–846 (2006)
9. Farenzena, M., Fusiello, A., Gherardi, R., Toldo, R.: Towards unsupervised reconstruction of architectural models. In: Proceedings of Vision, Modeling, and Visualization 2008, pp. 41–50 (2008)
10. Stiefelhagen, R., Finke, M., Yang, J., Waibel, A.: From gaze to focus of attention. In: Huijsmans, D.P., Smeulders, A.W.M. (eds.) VISUAL 1999. LNCS, vol. 1614, pp. 761–768. Springer, Heidelberg (1999)
11. Preparata, F.P., Shamos, M.I.: Computational Geometry. An Introduction