# Real Time Detection of Social Interactions in Surveillance Video

Paolo Rota, Nicola Conci, and Nicu Sebe

University of Trento, via Sommarive 5, Povo (TN) Italy

**Abstract.** In this paper we present a novel method to detect the presence of social interactions occurring in a surveillance scenario. The algorithm we propose complements motion features with proxemics cues, so as to link the human motion with the contextual and environmental information. The extracted features are analyzed through a multiclass SVM. Testing has been carried out distinguishing between casual and intentional interactions, where intentional events are further subdivided into normal and abnormal behaviors. The algorithm is validated on benchmark datasets, as well as on a new dataset specifically designed for interactions analysis.

## 1 Introduction

The research in video surveillance and environmental monitoring has revealed a recent trend in bringing the analysis of the scene to a higher level, shifting the attention from traditional topics, such as tracking and trajectory analysis [1], towards the semantic interpretation of the events occurring in the scene [2,3]. In particular, behavior analysis in terms of action and activity recognition has emerged as a relevant subject of research, especially for classification and anomaly detection purposes. Important contributions to the field have been proposed by Scovanner et al. [4], in which authors learn pedestrian parameters from video data to improve detection and tracking, and by Robertson et al. [5] where human behavior recognition is modeled as a stochastic sequence of actions described by trajectory information and local motion descriptors.

Bringing the analysis to a higher level of interpretation involves understanding the real social relationships undergoing between subjects, thus requiring to extend the analysis domain also to psychology and sociology. To this aim, the proxemics theory can be effectively exploited to observe the human relationships captured by a surveillance camera [6,7].

The goal of proxemics is to measure the social distance between subjects in order to infer interpersonal relationships. In this area, the works by Cristani et al. [8] aim at understanding the social relations among subjects when sharing a common space. The authors detect the so-called F-Formations present in the scene, thus inferring whether an interaction between two or more persons is occurring or not. A recent and relevant approach based on proxemics has been proposed by Zen et al. [9]. The authors identify proxemics cues in order

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 111–120, 2012.

to discriminate personality traits as *neuroticism* and *extraversion*, and use the collected data to construct the corresponding behavioral model. The acquired data is then used to improve the accuracy of the tracking algorithm. A similar approach has been proposed by Pellegrini et al. [10], using the social force model [11]. The solution proposed in [10] considers each subject as an agent, for which the model of motion has to be optimized, so as to prevent collisions with the other entities moving in the scene. The authors consider every agent as driven by its destination, taking into account, besides position, also additional parameters like velocity and direction of motion. The collected data is then used to build a model to measure the proximity level between the subjects, and to construct an avoidance function. A very recent approach is the work by Cui et al. [12]. The authors extract an *interaction energy potential* to model the relationships ongoing among groups of people. The relationship between the current state of the subject and the corresponding reaction is then used to model normal and abnormal behaviors. The authors also claim that their approach is independent from the adopted tool for human motion segmentation.

A hierarchical approach is instead proposed by [13] where human behavior is described at multiple levels of detail ranging from macro events to low-level actions. Authors exploit the fact that social roles and actions are interdependent one to each other and related to the macro event that is taking place.

In this work we define the interaction as a combination of energy functions that capture the state of a subject in the social context he moves. Since tracking is out of the scope of this work, our goal is to build a classifier to identify and recognize different types of behaviors. A novel aspect we introduce with respect to [10], consists in the insertion of an *intentionality* parameter in the processing chain, targeted at distinguishing between intentional and casual interactions. This term, provided by the proxemics information, is used to weight the interaction patterns acquired in real-time on a sliding window basis. The output of the function is then brought into the Fourier domain, thus removing the temporal correlation of the samples, and eventually fed into an SVM classifier. We have devised three different scenarios: (i) casual interaction, (ii) normal, and (iii) abnormal interaction. The interactions of type (i) refer to non-intentional events, while the type (ii) and (iii) reveal intentional interactions, divided into regular and potentially dangerous events.

The method has been tested on three datasets specifically chosen for human interaction analysis.

## 2 Methodology

According to the proxemics principles, distances can say a lot about the relationships going on between people, about their intimacy level, making it possible to distinguish between intentional and non-intentional behavioral cues. This information is generally variable in space in time and depends on the location in which a person stands, on the density of people in the area, but also on cultural and religious differences.

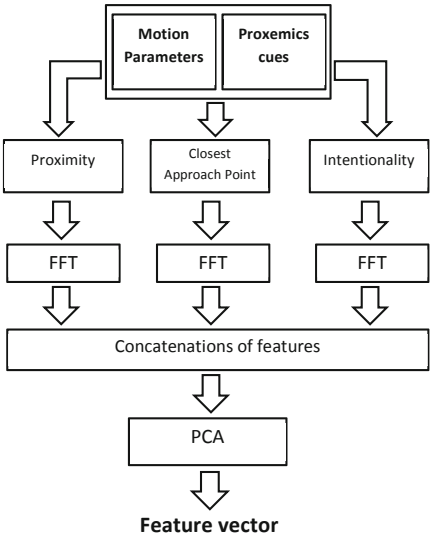



**Fig. 1.** Flowchart of the proposed architecture

Fig. 1 shows the proposed architecture for social interaction analysis, for which we will provide additional details in the next subsections.

### 2.1 Proxemics Parameters

In the model we propose, we follow the path covered by Pellegrini et al. [10] in order to capture the salient motion features that can be associated to an interaction. Each subject $i$ is modeled at each time $t$ by a state vector of parameters that takes into account the current position and velocity:

$$S_i(t) = [\mathbf{p}_i(t), \mathbf{v}_i(t)] \tag{1}$$

At each time instant $t$ it is then possible to model the distance between each pair of subjects $(i, j)$ as:

$$d_{ij}^2 = \left\| \mathbf{p}_i + t\mathbf{v}_i - \mathbf{p}_j - t\mathbf{v}_j \right\|^2 \tag{2}$$

By defining $\mathbf{k}_{ij}^t = \mathbf{p}_i^t - \mathbf{p}_j^t$ and $\mathbf{q}_{ij}^t = \mathbf{v}_i^t - \mathbf{v}_j^t$ and applying the derivative with respect to $t$ in Eq. (2), it is possible to find the time instant $t^*$ at which the distance $d_{ij}^*$ between the subjects is minimized.

$$t^* = -\frac{\mathbf{k} \cdot \mathbf{q}}{\|\mathbf{q}\|^2}, \quad d_{ij}^{*2} = \left\| \mathbf{k} - \frac{\mathbf{k} \cdot \mathbf{q}}{\|\mathbf{q}\|^2} \mathbf{q}^\top \right\|^2 \tag{3}$$

Eq. (3) is the estimate for the closest point (and the corresponding time instant), at which the subjects will most probably meet. However, this piece of information, although relevant to check whether there is chance for $i$ and $j$ to interact in the next future, does not necessarily include details about their interaction level. An estimate can be obtained by building an energy functional between subjects $i$ and $j$ by measuring the evolution of the proximity between them over time:

$$E_{ij}^{c} = e^{-\frac{d_{ij}^{*2}}{2\sigma_d^2}} \tag{4}$$

In Eq. (4) $\sigma_d$ controls the variance of the function in order to make it more or less responsive. The output of Eq. (4) can be seen as a *collision warning*, and represents the closest distance at which the two subjects will be, given the current motion parameters (position, velocity and direction of motion). This element is important because it can be used as a hint to predict the future developments of the interaction.

In line with the previous statement, we define an energy function to model the actual distance between subjects. This parameter is useful to obtain a proper modeling of the social behavior, since an interaction is more likely to happen when two persons are closer rather than when they are far apart from each other.

$$E_{ij}^{d} = e^{-\frac{\left\|\mathbf{k}_{ij}^{w}\right\|^2}{2\sigma_w^2}} \tag{5}$$

In [10], and for tracking purposes, the authors use the term $E_{ij}^{d}$ as a weight to model the outcome of Eq. (4) together with another term depending on the angle between the direction of motion of $i$ and the position of $j$. Our goal is however different, since we want to understand the dynamics of the interaction. Furthermore, the direction information, is in general noisy, particularly in the case of unrestricted video scenes, and for these reasons it has been discarded from out model.

In order to model the intentionality of an interaction, we adopt the so-called *O-space* [14]. The *O-space* consists of a circular area between the subjects, located in the direction of their gaze. It can be seen as the interaction space, namely the area comprised between two people interacting and facing one to each other.

By means of this definition, the *O-Space* can be used as a selectivity criterion, i.e. to inform about the presence of an interaction. The *O-Space* is in general defined as a static and non-deformable area right in front of the person and is not suitable for dynamic motion models, in which interactions can occur also in case the subjects move (e.g. walking together). Therefore, in our proposal we borrow the idea of the *O-space* as an area of attention of the subject, which can be adopted to infer the intentionality (or causality) of an interaction. In our model the *O-space* is positioned along the direction of motion of the subject and its center varies depending on his velocity. This gives us the opportunity of handling also dynamic interactions, and not only static events.

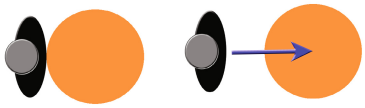
The position of the *O-space* is defined as:

$$\begin{aligned} Ox &= p_x + a_x \Lambda \sin(\theta) \\ Oy &= p_y - a_y \Lambda \cos(\theta) \end{aligned} \tag{6}$$

where $p_x$ and $p_y$ are the coordinates of the subject, $\Lambda$ is the displacement of the subject from the previous frame, $a_x$ and $a_y$ are tuning parameters depending on the field of view of the camera, and $\theta$ is the absolute direction of motion. The *O-space* area is used to calculate the intentionality component of the interaction, similarly to what we did for the proxemics information:

$$E_{ij}^{o} = e^{-\frac{\left\|\mathbf{k}_{ij}^{o}\right\|^2}{2\sigma_o^2}} \tag{7}$$

where $k_{ij}^{O}$ is the distance between the *O-space* centers of subject $i$ and $j$ , respectively. This parameter allows to filter out the noisy information collected by the other terms (for example two people very close but facing in opposite directions), thus reducing the chances of false positives returned in the presence of casual interactions of subjects standing nearby. The *O-space* model we have adopted is shown in Fig. 2.



(a) (b)

**Fig. 2.** O-space modeling. The figure represents the two cases in which the subject is (a) standing still, and (b) when he is moving from left to right. In the latter case the O-space shifts in the direction of motion proportionally with its velocity.

### 2.2 Feature Extraction

Following the flow chart in Fig. 1 we collect the proxemics values $E_{ij}^{d}(t), E_{ij}^{c}(t)$, $E_{ij}^{o}(t)$ in a given temporal window (128 samples in our examples), and at each time instant we apply the FFT (Fast Fourier Transform) (8) on the window samples. At this stage, the importance of the FFT is to eliminate the temporal correlation of the samples by only considering the contribution they bring into the interaction in terms of dynamics of that specific event.

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\imath 2\pi k \frac{n}{N}} \quad k = 0, ..., N-1 \tag{8}$$

The next step consists of concatenating the three sets of features to construct the feature vector that will be analyzed by the classifier. This process is carried out at every time instant, resulting in a large number of parameters (128x3).

Therefore, we apply a dimensionality reduction through Principal Component Analysis (PCA). Accordingly, the training set is arranged in a $n \times m$ matrix where $n$ is the number of samples and $m$ the number of features. From the matrix $X$ the eigenvalues of the related zero mean covariance matrix are extracted and the obtained vector is sorted by magnitude in descending order. The first value is the so-called principal component. From eigenvalues vector we can compute eigenvectors $m \times m$ matrix.

$$Y = W_s^T X \tag{9}$$

As shown in (9) the feature space has now been reduced, restraining the training set to a new matrix of size $n \times s$ where $s < m$ is the number of eigenvectors that we consider as relevant for our analysis.

Now that we have constructed our training set, we adopt a similar procedure for prediction. Each new incoming sample consists of a $1 \times m$ vector that is processed as $X$ in (9) obtaining as output a $1 \times s$ vector. This new vector is the input for the SVM, from which we will classify the type of the ongoing interaction.

### 2.3 Classification Procedure

After obtaining the reduced feature space, classification is computed using a kernel based SVM. Since the classification output strongly depends on the data used for training, let us briefly see what are the main steps we follow to obtain a reliable training set:

- Select the training videos representing the three classes that we want to classify with the frame-by-frame interaction labeling (manually done in a previous stage);
- Compute the interaction values as presented in Section 2.1 for the whole duration of the video;
- Segment the interaction values in accordance with the labels;
- Run the sliding window over the segmented interaction values, and consider each step as a feature vector;
- Transform each feature vector in the FFT domain and reduce the dimensionality using Principal Component Analysis;
- The resulting arrays consist of the features space for the classifier, which will be tuned by cross validation optimization to estimate the best configuration for the class separation.

It is worth noting that samples for training are picked randomly and in equal number for each class from the dataset, in order to avoid any possible bias in the training and to prevent overfitting of a particular class with respect to the others.

In the test phase the procedure simply consists of collecting the sliding window data at each time step, compute the FFT transform and the PCA decomposition using the training eigenvectors, thus building the new space. Data are then sent to the classifier for the final class prediction.

## 3 Results

*Datasets.* To validate our method we have used three different datasets: our own dataset SI (Social Interactions) Dataset [15], a selection of video sequences collection of YouTube CCTV videos (different contexts) and some sequences taken from the BEHAVE database.

The SI Dataset has been acquired to specifically address the topic of interactions analysis, since the number of social interactions occurring in more traditional datasets such as the PETS is limited, making it difficult to obtain sufficient statistical evidence. The set consists of 12 fully annotated video sequences of different length recorded at 25 FPS. Sequences mainly represent regular daily life behaviors such as people chatting, walking together or simply crossing each other. The dataset also includes more critical types of interactions, simulating

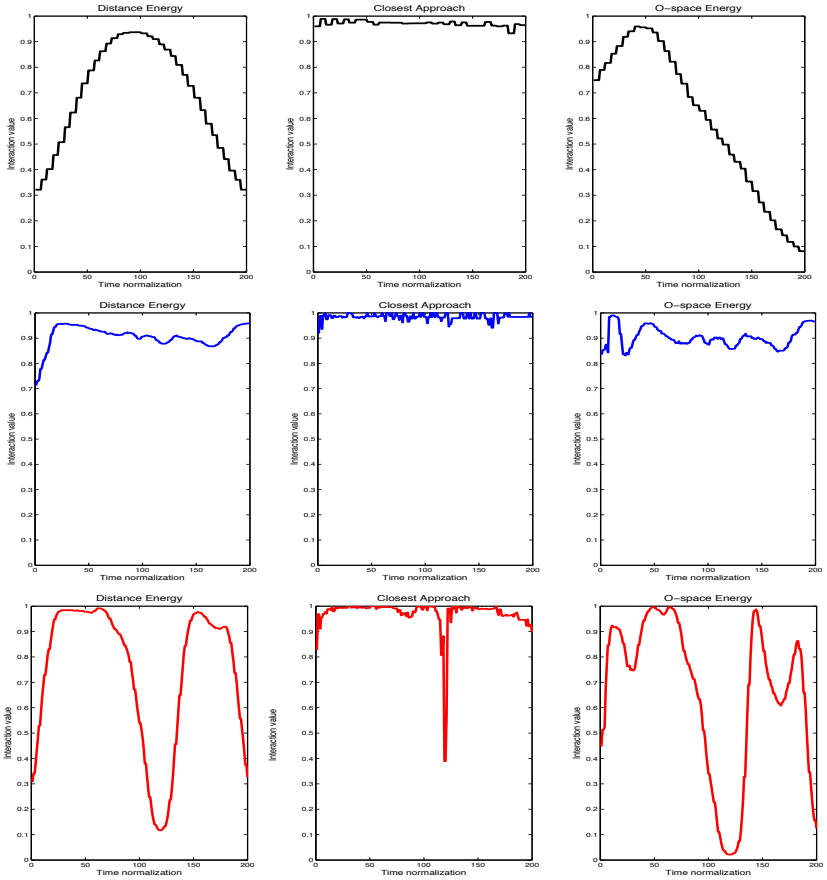



**Fig. 3.** Energy functions for distance (left), closest point of approach (center), and O-Space (right), in presence of two people crossing (first row), chatting (second row), and fighting (third row)

fights. The video sequences are recorded outdoor, under three different views, for which we will use here only the bird's eye view for similarity with the other datasets. For our experiments, and considering that tracking is out of the scope of this paper, we use the collected ground truth, from which it is possible to extrapolate all the necessary parameters required by our method.

The YouTube dataset is composed by 4 video sequences recorded in as many different locations. This dataset is not homogeneous because the videos come from different sources, with different view angles and different fields of view. For there reasons the videos are very challenging, since they represent real-life situations, and are not acquired with any specific purpose.

From the BEHAVE dataset [16] we have included in the experiments two different segments regarding different behavioral situations. Also here videos are acquired from far range, and are only partially annotated. We have then collected the corresponding ground truth.

*Experiments.* As mentioned in Section 2, classification is achieved via a multiclass SVM with Gaussian kernel. The number of training samples for each dataset is 1200, balanced over the three classes (400x3). In the training phase the best SVM parameters have been estimated by cross-validation. The testing phase takes as input the SVM parameters and the interaction parameters used to compute the interaction measure. These parameters are estimated through an exhaustive search and they differ in relation with the properties of the monitored area (range, field of view, angle). The proposed architecture allows computing

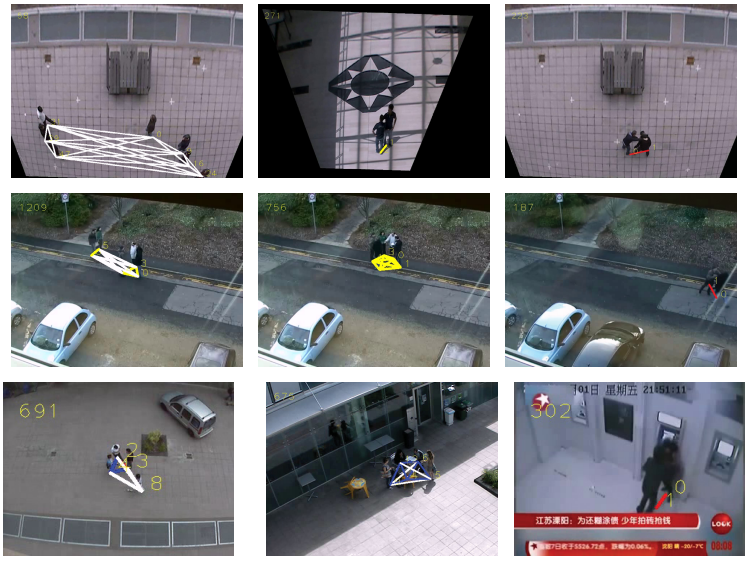



**Fig. 4.** Sample interactions taken from the three datasets. The first column indicates casual interactions, the central column refers to normal interactions, while the last column signals the presence of abnormal interactions.

the interaction measure on-line, without waiting for the end of the interaction. In fact, the complexity of the algorithm is negligible, compared to computational resources required for people detection and tracking. In Fig. 3 the energy functions obtained from three different sample sequences are shown.

In terms of numerical results we present two different tables, where it is possible to observe the effectiveness of our approach, especially in unconstrained scenarios, in which the interpretation of the interactions could be problematic. As it can be noticed from Table 1 and Table 2, the algorithm performs in general well especially in detecting the presence of an interaction, in all three datasets used for testing. As far as the class 3 is concerned (anomalous events) and considering the complexity of the task, the improvement given by the O-Space term is considerable (more than 20% in precision) due to the capability of better isolating the interacting subjects. A graphical presentation of the classification process is shown in Fig. 4. Here, each line reports three snapshots taken from the different datasets, each of them representing one of the classes. White lines (left column) indicate that no interaction is currently ongoing, yellow lines (center column) refer to normal interactions, while red lines (right column) indicate the presence of an abnormal event.

**Table 1.** Performance comparison of the proposed algorithm with and without the O-Space energy on the three datasets

| | | O-space Method | | | Without O-space Method | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | HitRate | Precision | Recall | HitRate |
| SI | Casual | 93,3% | 94,1% | 88,5% | 93,5% | 91,1% | 86,1% |
| | Normal | 75,1% | 76,1% | | 63,3% | 77,5% | |
| | Abnormal | 55,3% | 48,7% | | 55,4% | 45,3% | |
| Behave | Casual | 75,8% | 93,8% | 90,1% | 75,3% | 93,8% | 90,1% |
| | Normal | 98,0% | 93,9% | | 97,7% | 94,3% | |
| | Abnormal | 42,5% | 27,5% | | 43,6% | 22,6% | |
| YouTube | Casual | 88,2% | 90,2% | 82,7% | 66,3% | 76,2% | 60,1% |
| | Normal | 84,4% | 80,2% | | 70,9% | 43,7% | |
| | Abnormal | 38,2% | 40,3% | | 11,0% | 17,7% | |

**Table 2.** Confusion matrices for the three datasets obtained using the O-Space energy

| | | Casual | Normal | Abnormal |
|---|---|---|---|---|
| SI | Casual | 94,07% | 3,68% | 2,25% |
| | Normal | 18,49% | 76,17% | 5,34% |
| | Abnormal | 37,56% | 15,34% | 47,10% |
| Behave | Casual | 93,82% | 3,53% | 2,65% |
| | Normal | 3,98% | 93,92% | 2,10% |
| | Abnormal | 61,83% | 10,92% | 27,25% |
| YouTube | Casual | 90,16% | 5,00% | 4,85% |
| | Normal | 12,74% | 80,23% | 7,03% |
| | Abnormal | 29,71% | 29,92% | 40,37% |

## 4 Conclusion

In this paper we have proposed a tool to analyze social interactions in surveillance video, combining traditional metrics based on distance and velocity, and proxemics cues. Proxemics is handled as an intentionality parameter, giving the opportunity to better focus on the events of interest by considering only the moving subjects whose motion patterns demonstrate a will to interact. The method has been evaluated on three different datasets, confirming the viability of the method in recognizing different types of interactions. One of the datasets, specifically designed for social interactions analysis is provided by the authors as an additional contribution of the paper.

## References

1. Piotto, N., Conci, N., De Natale, F.: Syntactic matching of trajectories for ambient intelligence applications. IEEE Transactions on Multimedia 11(7), 1266–1275 (2009)
2. Zhang, Y., Ge, W., Chang, M., Liu, X.: Group context learning for event recognition. In: WACV (2010)
3. Turaga, P., Chellappa, R., Subrahmanian, V., Udrea, O.: Machine recognition of human activities: A survey. IEEE Transactions on Circuits and Systems for Video Technology 18(11), 1473–1488 (2008)
4. Scovanner, P., Tappen, M.: Learning pedestrian dynamics from the real world. In: ICCV, pp. 381–388 (2009)
5. Robertson, N., Reid, I.: Behaviour understanding in video: a combined method. In: ICCV, vol. 1 (2005)
6. Hall, E.: The hidden dimension, vol. 6. Doubleday, New York (1966)
7. Hall, E.: The silent language. Anchor (1973)
8. Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Del Bue, A., Menegaz, G., Murino, V.: Social interaction discovery by statistical analysis of f-formations. In: Proceedings of British Machine Vision Conference (2011)
9. Zen, G., Lepri, B., Ricci, E., Lanz, O.: Space speaks: towards socially and personality aware visual surveillance. In: MPVA 2010, pp. 37–42. ACM (2010)
10. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: ICCV (2009)
11. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: CVPR (2009)
12. Cui, X., Liu, Q., Gao, M., Metaxas, D.: Abnormal detection using interaction energy potentials. In: CVPR, pp. 3161–3167 (2011)
13. Lan, T., Sigal, L., Mori, G.: Social roles in hierarchical models for human activity recognition. In: CVPR (2012)
14. Cristani, M., Paggetti, G., Vinciarelli, A., Bazzani, L., Menegaz, G., Murino, V.: Towards computational proxemics: Inferring social relations from interpersonal distances. In: SocialCom, pp. 290–297 (2011)
15. Rota, P., Zhang, B., Ullah, H., Conci, N.: Unitn social interactions dataset. University of Trento, Italy (2012), `http://mmlab.science.unitn.it/USID/`
16. Laghaee, A.: Behave dataset (2007), `http://homepages.inf.ed.ac.uk/rbf/BEHAVE/`