Exploiting visual search theory to infer social interactions

Paolo Rota[a], Duc-Tien Dang-Nguyen[a], Nicola Conci[a], and Nicu Sebe

[a]University of Trento, Via Sommarive 5, 38123 Trento, Italy;

ABSTRACT

In this paper we propose a new method to infer human social interactions using typical techniques adopted in literature for visual search and information retrieval. The main piece of information we use to discriminate among different types of interactions is provided by proxemics cues acquired by a tracker, and used to distinguish between intentional and casual interactions. The proxemics information has been acquired through the analysis of two different metrics: on the one hand we observe the current distance between subjects, and on the other hand we measure the O-space synergy between subjects. The obtained values are taken at every time step over a temporal sliding window, and processed in the Discrete Fourier Transform (DFT) domain. The features are eventually merged into an unique array, and clustered using the K-means algorithm. The clusters are reorganized using a second larger temporal window into a Bag Of Words framework, so as to build the feature vector that will feed the SVM classifier.

Keywords: Social Interactions, Proxemics, Human Behavior, Action Recognition

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],[2] towards the semantic interpretation of the events occurring in the scene.[3,4] 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.,[5] in which authors learn pedestrian parameters from video data to improve detection and tracking; Robertson et al.[6] model human behaviors 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 social relationships undergoing between subjects, thus requiring an extension of the analysis domain also including psychology and sociology. To this aim, the proxemics theory can be effectively exploited to observe the human relationships when captured by a surveillance camera.[7,8] According to Hall's theory, each person speaks a *silent language*, related to the behavior of the subject, and expressed in terms of motion and body pose. The main proposition of his studies consists of the correlation between the distance among people and the corresponding relationship ongoing between them. The interpersonal distance can be modeled as:

- *Intimate distance:* between $0 - 45$ cm;
- *Personal distance:* between $45 - 120$ cm, for friendship relationship;
- *Social distance:* between $120 - 350$ cm for formal relationship;
- *Public distance:* over 350 cm for public relationships.

Following similar principles, Cristani et al.[9] aim at understanding the social relations among subjects when sharing a common space. The authors detect the so-called F-Formations, in order to infer the presence of an ongoing interaction between two or more persons. An approach based on proxemics is proposed by Zen et al.[10] The authors identify proxemics cues in order 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.,[11] using the social force model.[12] The solution proposed in[11] 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 model the proximity level between subjects, in order to construct an avoidance function. Cui et al.[13] 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 proposed by Lan et al.[14] 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.

The goal of this paper is to recognize different types of social interactions, approaching the problem from a slightly different point of view, extending our previous work[15] and relying exclusively on proxemics cues.

Interactions are defined as a combination of energy functions that capture the state of a subject in the social context he moves. Considering that the main goal of this work is to construct a classifier to recognize different types of interactions, the details related to the tracking algorithm will not be discussed.

Comparing to the existing state of the art,[11] we propose to insert 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 by applying a DFT (thus removing the temporal correlation of the samples), and then clustered using K-means. At this stage we collect another sliding window of clusters to fill a Bag of Words array that will be classified using 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 two subjects, about their intimacy level, making it possible to distinguish between intentional and non-intentional behaviors. This information generally depends on space and time, but also on the location in which a person stands, on the density of people in the area, on cultural and religious differences.

The overall architecture we propose in this paper to analyze social interactions is shown in Fig. 1. Additional details about the method are provided in the next subsections.

2.1 Proxemics parameters

As we will see in the next paragraphs, in our model we propose to use two different metrics in order to capture the salient motion features that can indicate the presence of an interaction.

Each subject i is associated at each time t with 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:

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

From (2) we define an energy function that models the actual distance between subjects (3), since an interaction is more likely to happen when two persons are closer rather than when they are far apart from each other.

$$E_{ij}^d = e^{-\frac{\left\|\mathbf{d}_{ij}^d\right\|^2}{2\sigma_d^2}} \tag{3}$$

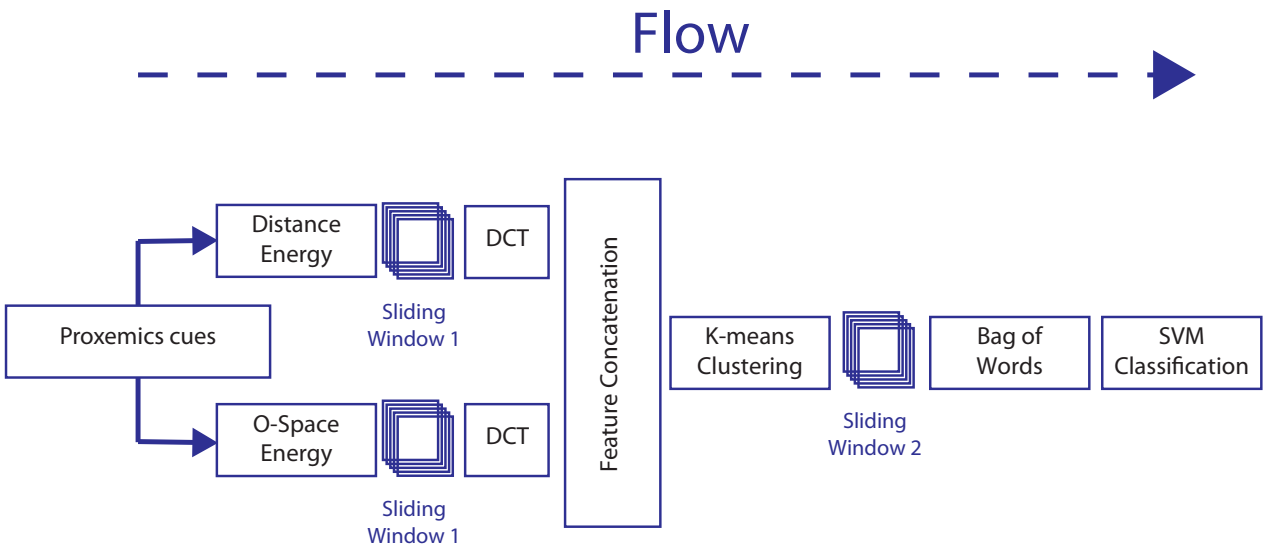


Figure 1. Flowchart of the proposed architecture.

The value of σ_d is related to the responsiveness of the function and it can vary depending on the camera system parameters.

In order to model the intentionality of an interaction, we adopt the so-called *O-space*.[16] 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, represented by the area comprised between two people interacting and facing one to each other.

According to 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 in front of the person and it is not suitable for the inclusion in 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 the velocity. This gives us the opportunity of handling also dynamic interactions, and not only static events.

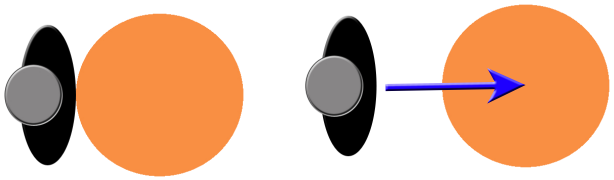
The position of the *O-space* is defined as:

$$\begin{aligned} Ox &= p_x + a_x \Lambda \sin(\theta) \\ Oy &= p_y - a_y \Lambda \cos(\theta) \end{aligned} \tag{4}$$

where p_x and p_y are the coordinates of the subject, Λ 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 θ 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:

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

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 occurring in presence of casual interactions of subjects standing nearby. The *O-space* model we have adopted is shown in Fig. 2.



(a) (b)

Figure 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 figure shows how the O-space shifts in the direction of motion proportionally with its velocity.

2.2 Feature extraction

In accordance with the flowchart proposed in Fig. 1, the proxemics values $E^d_{ij}(t)$ and $E^o_{ij}(t)$ are collected over a temporal window of fixed length N. At each time instant we apply the DFT (6) on the window samples. The role of the DFT is here to reduce the temporal correlation of the samples by only considering the contribution they bring into the interaction in terms of dynamics for each specific event.

$$v_t = \sum_{n=0}^{N-1} x_n e^{-\imath 2\pi t \frac{n}{N}} \qquad t = 0, ..., N-1 \tag{6}$$

The next step consists of clustering the features: each temporal window is represented as a vector, denoted as $\mathbf{v_j} = [v_1, v_2, \ldots, v_{2N}]$, which is the result of the concatenation of the two transformed energy functions. Then, all vectors are clustered into K clusters by using the traditional K-means algorithm.[17] Given a set of vectors, computed as mentioned above, K-means algorithm clusters each $\mathbf{v_j}$ into $S = \{s_1, s_2, ...s_K\}$, by minimizing the sum of distances:

$$\arg\min_{S} \sum_{i=1}^{K} \sum_{\mathbf{v_j} \in s_i} ||\mathbf{v_j} - \mathbf{u_i}||^2 \tag{7}$$

where $\mathbf{u_i}$ is the mean of s_i and $||\cdot||$ is the L_2 distance. The traditional K-means algorithm solves Eq. 7 by starting from an initial configuration of the centers, which is in our case randomly initialized by taking K vectors from the initial set. Each vector $\mathbf{v_j}$ is assigned to the closest cluster. The centers are then updated by computing the new mean of the clusters:

$$\mathbf{u_i} = \frac{1}{|s_i|} \sum_{\mathbf{v_j} \in s_i} \mathbf{v_j} \tag{8}$$

At each frame each pair of subjects is described by a center s_i where $i = 1, \ldots, k$. A second temporal window of length M collects the values $\mathbf{f} = [s_1, \ldots, s_M]$. In order to create the final set of features, a Bag of Words is computed over the vector $\mathbf{f}$. This method incorporates temporal and spatial information about the two subjects into the feature vector, moreover it generalizes the time information with a double layer created by the two temporal windows.

2.3 Classification procedure

After obtaining the new feature space, features classification is computed using a kernel-based SVM. Since the classification output strongly depends on the data used for training, let us briefly recap the main steps we follow to obtain a reliable training set (see also Fig. 1) :

- Select the training videos representing the three classes and label the type of interaction on a frame-by-frame basis;
- Compute the interaction values as presented in Section 2.1 for the entire duration of the video;

- Run the sliding window over the segmented interaction values, creating a preliminary feature set;
- Transform each feature vector in the Fourier domain using DFT;
- Each vector can now be assigned to the corresponding center, as obtained by the K-means algorithm, thus creating a different set of vectors;
- A second sliding window scans the new data and fills a Bag of Words, creating a new set of features, which dimension equals the number of clusters K;
- Classification is then performed using SVM.

3. RESULTS

Datasets. For the validation of the proposed method we have considered three different datasets: our own dataset (defined as SI - Social Interactions dataset),[15] a selection of video sequences collected on YouTube from CCTV videos (different contexts) *, and a subset of the sequences of the BEHAVE dataset.[18]

The SI dataset has been acquired to specifically address the topic of interactions analysis in surveillance contexts. Therefore, we provide a brief explanation of its content. The set consists of 12 fully annotated video sequences of different length recorded at 25 FPS. Video sequences mainly represent regular daily life behaviors such as people chatting, walking together or simply crossing each other. The dataset also includes more complex types of interactions, as the simulation of fights. Videos 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. In our experiments, and considering that tracking is out of the scope of this paper, we process 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 fields of view. For these reasons the videos are very challenging, since they represent real-life situations, and are not acquired with any specific purpose.

From the BEHAVE dataset we have included in the experiments two segments regarding different types of interactions. Also in this case, videos are acquired from far range, and are only partially annotated. We have then collected the corresponding ground truth.

Experiments. The experiments consist in comparing our approach based on K-Means and Bag Of Words, against the simple classification applied on the interaction metrics and processed by a standard SVM classifier.

In our experiment, each vector $\mathbf{v_j}$ has the size of 200 elements (we use two different features representing distance and O-space, respectively, on temporal windows of 100 frames each). All the parameters are computed through an exhaustive search performed over all the three datasets. Those values are highly connected with the dataset characteristics (i.e. field and angle of view, etc...), nevertheless we assumed them constant for each set of data to simplify the results comparability. Parameters are reported in Table 1:

Table 1. Parameters configuration used in the experiments.

Parameter	**Value**
σ_d	100
σ_o	20
Clusters (K-means)	15
Interaction window length	100
BOW window	80

As mentioned in Section 2, classification is achieved via a multi class SVM with Gaussian kernel. The overall 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 results are presented in terms of hit-rate, and the corresponding confusion matrices are shown in Table 2.

*http://mmlab.science.unitn.it/USID/

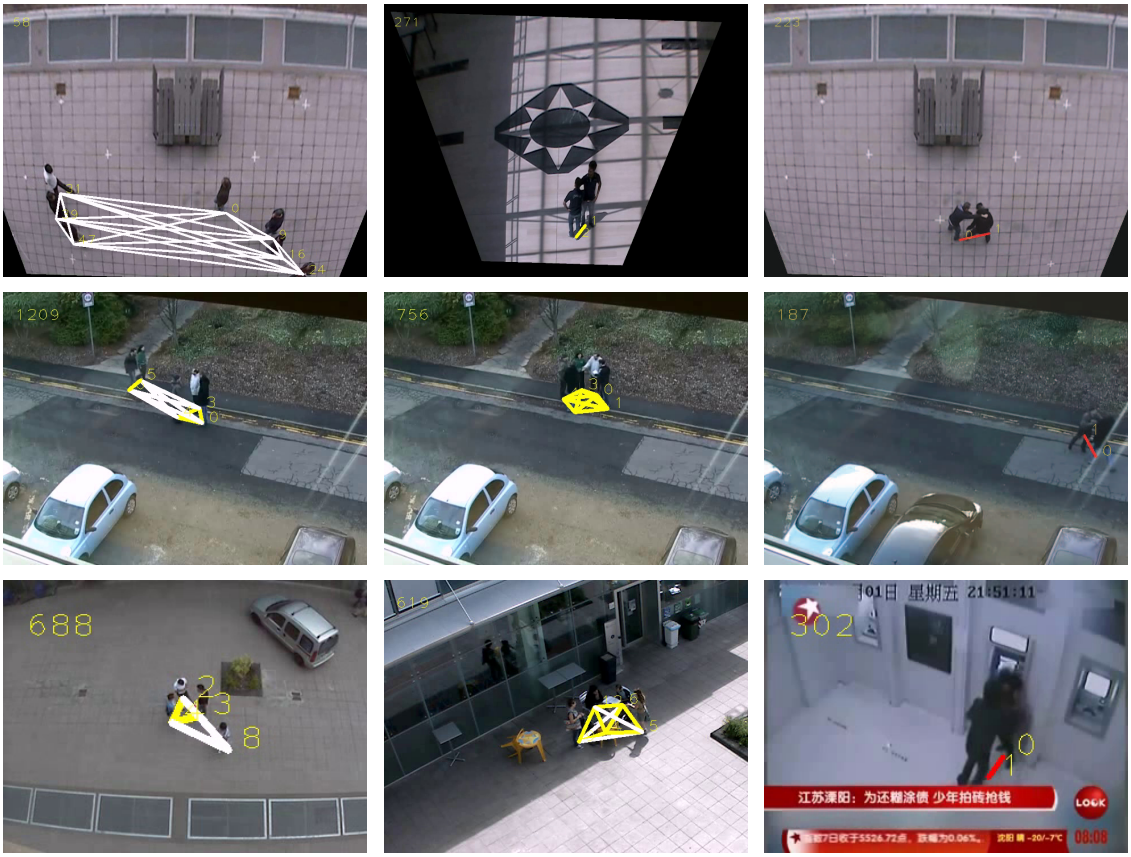


Figure 3. Sample interactions taken from the three datasets: casual interactions (left column), normal interactions (center), abnormal interactions (right).

Table 2. Performance comparison of the proposed approach against the direct SVM approach.

		Our Method			**Direct SVM**		
		No Interaction	Normal	Abnormal	No Interaction	Normal	Abnormal
Behave	No Interaction	70,50%	9,73%	19,78%	51,34%	3,75%	44,91%
	Normal	3,48%	75,29%	21,23%	2,08%	75,23%	22,69%
	Abnormal	8,82%	11,91%	79,26%	14,96%	17,94%	67,11%
SI	No Interaction	94,59%	4,43%	0,98%	89,23%	7,63%	3,13%
	Normal	19,91%	78,56%	1,53%	0,00%	63,87%	36,13%
	Abnormal	17,66%	26,05%	56,29%	7,09%	30,26%	62,65%
YouTube	No Interaction	63,34%	13,32%	23,34%	59,49%	5,29%	35,22%
	Normal	7,43%	87,58%	4,99%	6,22%	61,47%	32,31%
	Abnormal	30,33%	3,95%	65,72%	16,32%	15,03%	68,65%

From the figures in Table 2 we can observe a general increment of performances of our method compared to the standard SVM. The Bag of Words model allows for a stronger correlation between adjacent frames. In fact, the feature vectors of two adjacent frames of the same pair of subjects only differ of no more than two values, thus resulting a more robust classification against fast changes in the short period.

A graphical presentation of the classification process is shown in Fig. 3. 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.

4. CONCLUSION

In this paper we have proposed a method to analyze social interactions in surveillance video, combining traditional metrics based on different proxemics cues combined in a Bag Of Words framework. Proxemics is handled by position, velocity and an intentionality parameter; this allows to better focus on the events of interest, by only considering the moving subjects, whose motion patterns are showing compatibility among each other. The approach proposes a real time classification of three types of interactions using a combination of algorithms as K-means, Bag of Words, and SVM classification. The method has been validated on three different datasets, confirming a general improvement of performance compared to a standard SVM classification.

REFERENCES

[1] Piotto, N., Conci, N., and De Natale, F., "Syntactic matching of trajectories for ambient intelligence applications," *Multimedia, IEEE Transactions on* **11**(7), 1266–1275 (2009).

[2] Broilo, M., Piotto, N., Boato, G., Conci, N., and De Natale, F., "Object trajectory analysis in video indexing and retrieval applications," *Video Search and Mining* , 3–32 (2010).

[3] Zhang, Y., Ge, W., Chang, M., and Liu, X., "Group context learning for event recognition," in [*WACV*], (2010).

[4] Turaga, P., Chellappa, R., Subrahmanian, V., and Udrea, O., "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video Technology* **18**(11), 1473–1488 (2008).

[5] Scovanner, P. and Tappen, M., "Learning pedestrian dynamics from the real world," in [*ICCV*], 381–388 (2009).

[6] Robertson, N. and Reid, I., "Behaviour understanding in video: a combined method," in [*ICCV*], **1** (2005).

[7] Hall, E., [*The hidden dimension*], vol. 6, Doubleday New York (1966).

[8] Hall, E., [*The silent language.*], Anchor (1973).

[9] Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Del Bue, A., Menegaz, G., and Murino, V., "Social interaction discovery by statistical analysis of f-formations," in [*Proceedings of British Machine Vision Conference*], (2011).

[10] Zen, G., Lepri, B., Ricci, E., and Lanz, O., "Space speaks: towards socially and personality aware visual surveillance," in [*MPVA'10*], 37–42, ACM (2010).

[11] Pellegrini, S., Ess, A., Schindler, K., and Van Gool, L., "You'll never walk alone: Modeling social behavior for multi-target tracking," in [*ICCV*], (2009).

[12] Mehran, R., Oyama, A., and Shah, M., "Abnormal crowd behavior detection using social force model," in [*CVPR*], (2009).

[13] Cui, X., Liu, Q., Gao, M., and Metaxas, D., "Abnormal detection using interaction energy potentials," in [*CVPR*], 3161–3167 (2011).

[14] Lan, T., Sigal, L., and Mori, G., "Social roles in hierarchical models for human activity recognition," in [*CVPR*], (2012).

[15] Rota, P., Conci, N., and Sebe, N., "Real time detection of social interactions in surveillance video," in [*Computer Vision–ECCV 2012. Workshops and Demonstrations*], 111–120, Springer (2012).

[16] Cristani, M., Paggetti, G., Vinciarelli, A., Bazzani, L., Menegaz, G., and Murino, V., "Towards computational proxemics: Inferring social relations from interpersonal distances," in [*SocialCom*], 290–297 (2011).

[17] MacKay, D., "An example inference task: Clustering," *Information Theory, Inference and Learning Algorithms* , 284–292 (2003).

[18] Laghaee, A., "Behave dataset." http://homepages.inf.ed.ac.uk/rbf/BEHAVE/ (2007).