High Level Event Detection based on Spatial Occupancy and Interpersonal Relationships

John Soldera, José Bins, Marcelo Cohen, Julio C. S. Jacques Junior, Soraia R. Musse, and Cláudio R. Jung

***Abstract*—This paper describes a new approach for event detection in video sequences. A tracking algorithm for oblique camera setups is initially used to extract trajectories in a training period, and a map of spatial occupancy of the scene is built. In the test stage, Voronoi Diagrams are used to obtain some information regarding interpersonal relationships, such as distances from neighbors, formation and classification of groups. A variety of complex events can be detected through a query formulated by the user, that may combine concurrent or sequential occurrences of simpler events based on either spatial occupancy or interpersonal relationships (e.g. group formation in a region with small spatial occupancy). These queries can be used to detect events on-the-fly as the video is processed, or applied to stored video databases.**

***Index Terms*—People tracking, event detection, spatial occupancy, interpersonal relationships.**

I. Introduction

With the decrease in price and increase in quality of video acquisition systems, the analysis of human motion from video sequences has become an important topic of research in the computer vision and pattern recognition communities, with several applications [1]–[3]. Among these applications, automatic or semi-automatic algorithms for video surveillance have gained increased attention in the past year, aiming to prevent criminal actions or terrorist actions.

In general, a suspect behavior can be characterized by several different aspects, such as motion (spatial occupancy), interaction with other people and the environment, gait analysis and gesture analysis, among others. This paper focuses on event detection based on two main aspects: spatial occupancy and interpersonal relationships. In a training period, people captured by a static camera are tracked using computer vision algorithms, and they are expected to follow designated pedestrian paths in structured environments. Based on the extracted trajectories, a Spatial Occupancy Map (SpOM) [4] of the scene is obtained, and it provides the location of expected flows of people. In the evaluation stage, we explore the temporal evolution of Voronoi Diagrams (VDs) to extract aspects related to interpersonal relationships, such as distance from neighbors across time, group formation and classification [5]. Simple events regarding either the spatial occupancy or interpersonal relationships can be easily detected, such as the detection of people walking on an unoccupied portion of space, or the formation of a voluntary group. The user can then use a specific grammar that allows the combination of these simple events to formulate more complex queries, and a finite automaton is used to match occurrences of these queries in video sequences. Such match can be performed on-the-fly, as the video is executed, or in stored databases.

The remainder of this paper is organized as follows. Section II presents some related works concerning event detection, while the proposed approach is presented in Section III. Some experimental results are provided in Section IV, Finally, the conclusions are presented in Section V.

II. Related Work

Several authors have presented methods for understanding the motion/behavior of people filmed by static cameras. There is a great variety of approaches and applications, ranging from global (crowd behavior) to local analysis (tracking of individual trajectories or body parts). Within this range of applications, there are methods that explore the coherence of tracked trajectories for unusual behavior detection, and others that rely on people interactions for motion understanding. Some of these techniques are briefly revised next (a more comprehensive review on human motion understanding and surveillance can be found in the survey papers [3], [6], [7]).

A. Motion Analysis

Part of existing method focus on trajectory coherence. Junejo et al. [8] proposed a method for detecting nonconforming trajectories of objects as they pass through a scene by comparing spatial similarity, velocity characteristics of trajectories and curvature features.

Fuentes and Velastin [9], [10] proposed event detection algorithms based on trajectories and foreground blobs designed for closed-circuit television (CCTV) surveillance systems. In their approach, some pre-defined events involving two or more persons can be detected (such as fights, attacks and vandalism), but the concept of grouping was not explicitly used.

Weiming et al. [11] presented a system for automatically learning motion patterns for anomaly detection and behavior

The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vanessa Testoni.

John Soldera, and Cláudio Jung are with the Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre-RS, Brazil (emails: {jsoldera, crjung}@inf.ufrgs.br).

José Bins is with Universidade Federal do Pampa - UNIPAMPA, Alegrete-RS, Brazil (email: josebins@unipampa.edu.br).

Marcelo Cohenn, Julio Jacques Junior and Soraia Musse are with Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS, Porto Alegre-RS, Brazil (emails: {marcelo.cohen, soraia.musse}@pucrs.br and julio.jacques@acad.pucrs.br).

This work was partially supported by CNPq.

Julio C. S. Jacques Junior thanks FAPERGS/CAPES for the support.

This paper is part of the Special Issue on Vision-based Human Activity Recognition.

Digital Object Identifier 10.14209/jcis.2015.8

prediction. To learn motion patterns, trajectories are clustered hierarchically using spatial and temporal information and then each motion pattern is represented with a chain of Gaussian distributions. In a similar approach [12], trajectories are clustered hierarchically using spatial and temporal information, to learn activity models. The proposed retrieval framework supports various queries including queries by keywords, multiple object queries, and queries by sketch.

Jung et al. [13] used motion displacement vectors to characterize trajectories, and proposed a clustering approach based on mixtures of Gaussians (where each component of the mixture represents one cluster). Their approach also includes a 4D histogram that models the position and the local velocity of each cluster, allowing the detection of transition points between clusters, bifurcation points and confluence points. The group of Wang [14] proposed a technique for trajectory analysis that simultaneously learns activities and semantic regions[1], which are jointly modeled using Dual Hierarchical Dirichlet Processes. With the detection of the semantic regions, other information can be obtained such as entry and exist points.

Morris and Trivedi [15] proposed a framework for live video analysis in which the behaviors of surveillance subjects are described using a vocabulary learned from recurrent motion patterns, for real-time characterization and prediction of future activities, as well as the detection of abnormalities. A three-stage process was used: learning interesting nodes by Gaussian Mixture Modeling, connecting routes using trajectory clustering, and encoding spatio-temporal activities using Hidden Markov Models (HMMs). Hu and colleagues [16] presented an incremental DPMM (Dirichlet Process Mixture Model) to cluster, model and retrieve trajectories. Each trajectory extracted by trackers is represented in the frequency domain, and clustered using an incremental DPMM that learns the number of clusters and can be updated on-the-fly (temporal changes in each trajectory can be detected as well, by using smaller tracks to build each trajectory). Also, a sketch-based scheme can be used to retrieve trajectories stored in a database based on similarity.

B. Human interactions

Another class of approaches focus on people interactions. Buxton and Gong [17] described a method to determine individual people behavior using Bayesian networks. In their approach, the objects dynamics are tracked and their behaviors are described as Bayesian networks, which contain information such as time and events. To classify events, the system classifies agents proximity as: not near, nearby, close, very close and touching.

Hosie and collaborators [18] proposed a method for group behavior detection that relies on *pair primitives*, which are pre-defined movements that can occur between two targets in the scene over one time sample. Oliver et al. [19] explored Coupled Hidden Markov Models (CHMMs) to model and recognize human tasks. Du and collaborators [20] proposed a similar approach using Dynamic Bayesian Networks (DBNs) instead of CHMMs. These approaches are able to detect some kinds of pre-defined human interactions (such as *follow*, or *approach + talk + continue together*), but are also limited mostly to interactions between two persons only.

Gong and Xiang [21], [22] explored Dynamic Probabilistic Networks (DPNs) for modeling temporal relationships among a set of different object events in the scene for a coherent and robust scene-level behavior interpretation. Although grouping is embedded in their approach, the main focus is activity recognition. Also, psychological aspects are not considered for grouping purposes. Wang and collaborators [23] proposed an approach for unsupervised activity perception in crowded scenes. Their method models atomic activities as low-level features, and multiagent interactions are modeled as distributions over atomic activities, using hierarchical Bayesian models based on Dirichlet processes. The whole model is based on motion cues, and not individual tracking.

Liu and Chua [24] presented a new method for modeling and classifying multi-agent activities based on observation decomposed hidden Markov models (ODHMMs), also proposed by the authors. In their approach, pre-defined activities with interacting people can be trained, using the relative distances between any two people as feature vectors. Despite the good results achieved for detecting "Snatch Thefts", the proposed method presents a high computational cost as the number of agents increases. Furthermore, tests were performed using manual tracking of people, making its practical application difficult to evaluate.

Wang and collaborators [23] proposed an approach for unsupervised activity perception in crowded scenes. Their method models atomic activities as low-level features, and multiagent interactions are modeled as distributions over atomic activities, using hierarchical Bayesian models based on Dirichlet processes. The whole model is based on motion cues, and not individual tracking. Although such approach may be adequate for crowded scenes (where individual tracking is very difficult), the lack of temporal coherence of motion cues can inflict in some errors.

Ryoo and Aggarwal [25] presented a methodology for automated recognition of complex human activities, recognizing high-level human actions and human-human interactions. In their approach, the user encodes the structure of a high-level human activity as a formal representation using a context-free grammar, and human activities are recognized by semantically matching constructed representations with actual observations. Their methodology allows the representation and recognition of complex human activities with a high recognition rate, but limited to two-person interactions. Ge et al. [26] proposed a method to discover pedestrian groups in a video sequence, using trajectories that projected into the ground plane and hierarchical clustering approach to identify and merge/split small groups of people. Cheng et al. [27] represented the problem of group activity recognition by a three-layered approach that gathers information about the individuals performing the actions, the possible pairs between two people and small groups.

Despite the existence of several approaches for event de-

[1]In [14], semantic regions are the intersections of paths commonly taken by objects.

tection based on trajectory coherence or simple interactions among people (mostly, between two persons), as far as we know there is no method that combines these two criteria in an efficient manner. This paper proposes a new method that explores Spatial Occupancy Maps (SpOMs) and interpersonal relationships to detect a broad variety of events, that can be formulated through queries using a grammar that contains simpler individual events. The proposed method is described next.

III. The Proposed Approach

A. People Tracking

The first step for automatic event detection is to obtain the trajectory of each person captured by the camera (tracking). Even though there are many approaches for people tracking [28]–[36], there is no gold standard that works well in all situations, with particular problems in the presence of shadows and people walking together. Although the focus of this paper is not on tracking itself, a new approach for people tracking considering specific camera setups (position and orientation of the camera) was developed, and it is briefly presented next.

1) Background Subtraction: As in most tracking algorithms with static cameras, the initial step of our algorithm is background subtraction. We adopted a simple and fast approach that includes shadow/highlight removal [37] to extract foreground blobs. In summary, the temporal median of each pixels is adopted as the background model, and the standard deviation of each pixel across time is computed. The distribution of the standard deviations for all pixels is used to obtain an estimate of the camera noise, and local spatial coherence is explored for foreground/background discrimination. Then, an approach for shadow and highlight detection based on pixel ratios is applied to remove foreground blobs generated by illumination changes.

The adopted background subtraction procedure can indeed prevent light shadows from being detected as foreground pixels, but strong (dark) shadows are still wrongly classified as foreground pixels, which may produce persons with very large blobs or connect different persons into the same blob. To identify each individual person from extracted blobs, the expected vertical position of a standing person in the projected image is searched, as explained next. With the proposed approach, cast shadows that are not aligned with the expected body orientation in image coordinates can be effectively discarded.

2) Camera Calibration: In this work, we assume that people walk on an approximately flat region. Given a set of 3D world coordinates (x, y, z), the mapping[2] to image coordinates (u, v) is given by a function $(u, v) = \boldsymbol{f}_i(x, y, z)$ given by [38]:

$$\begin{aligned} u &= \frac{c_1 x + c_2 y + k_1 z + c_3}{c_7 x + c_8 y + k_3 z + 1}, \\ v &= \frac{c_4 x + c_5 y + k_2 z + c_6}{c_7 x + c_8 y + k_3 z + 1}. \end{aligned} \quad (1)$$

Fig. 1 illustrates the coordinate axis uv (image) and the xy axis (world) adopted in this work. The z axis is orthogonal to the xy plane.

[2]In this work, radial distortion was not considered.

Fig. 1: Image axis in image (uv) and world (xy) coordinates for our environment.

To obtain the parameters c_i, $i = 1, ..., 8$, we first use the plane-to-plane mapping at $z = 0$ (i.e., the ground). We select $N_c \geq 4$ points on the ground, measure their coordinates in both world coordinates $(x, y, 0)$ and image coordinates (u, v), and solve Equation (1) for c_i. If $N_c > 4$, an overdetermined system arises, and it is solved by minimum squares. In fact, it is advisable to use $N_c > 4$, to account for measurement errors (we used $N_c = 6$ in our calibration).

Once the ground plane is calibrated, it is possible to use the expected height and geometry of standing people to estimate the values for k_i, $i = 1, 2, 3$. For that, a set of frames containing people is selected, and $N_k \geq 2$ persons distributed in different positions of the ground plane are selected. The position of the feet corresponds to a point (u_f, v_f) in image coordinates, that relates to a position $(x_f, y_f, 0)$ in world coordinates. If a certain position (u, v) in image coordinates is known to be in a certain height z in world coordinates, then it is possible to solve Equation (1) for (x, y), so that $(x, y) = \boldsymbol{f}_w(u, v, z)$. In particular, when $z = 0$, the function $\boldsymbol{f}_w(u, v, 0)$ is given by

$$\begin{aligned} x &= \frac{(c_8 c_6 - c_5)u + (c_2 - c_3 c_8)v + c_3 c_5 - c_2 c_6}{(c_5 c_7 - c_4 c_8)u + (c_1 c_8 - c_2 c_7)v + c_2 c_4 - c_1 c_5}, \\ y &= \frac{(c_4 - c_6 c_7)u + (c_3 c_7 - c_1)v - c_4 c_3 + c_1 c_6}{(c_5 c_7 - c_4 c_8)u + (c_1 c_8 - c_2 c_7)v + c_2 c_4 - c_1 c_5} \end{aligned} \quad (2)$$

and then (x_f, y_f) can be computed from (u_f, v_f).

The position of the head (in pixel coordinates) is given by (u_h, v_h), and assuming that the person is standing straight, it corresponds to a position (x_f, y_f, h) in world coordinates, where h is the height of the person. To simplify the procedure, we assume that all persons have an average height $h_m = 1.7$m, and then solve Equation (1) for k_i, using $(u, v) = (u_h, v_h)$, $(x, y, z) = (x_f, y_f, h_m)$. Again, to cope with measurement errors, it is advisable to use $N_k > 2$ (we used $N_k = 10$), and then solve Equation (1) by minimum squares.

It is important to note that this calibration procedure is manual, but it is performed only once (assuming that the camera is stationary). The measurement of control points on the ground in world coordinates must be carried out on site,

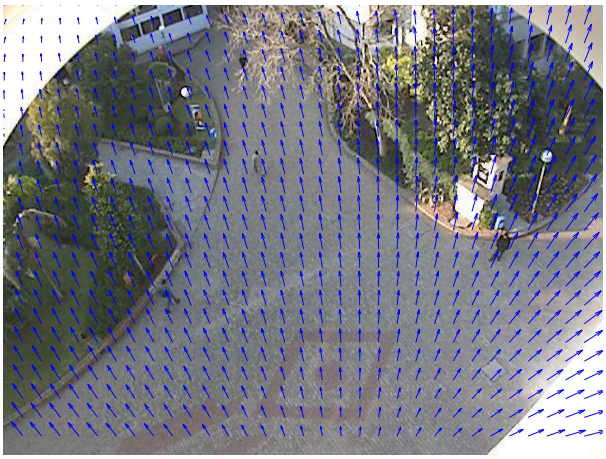


Fig. 2: Orientation field $\boldsymbol{o}(u, v)$ used to detect standing people for tracking.

and the procedure to compute k_i can be achieved by mouse clicking on the feet and head of selected subjects.

3) Head estimation: One way to track each individual person across time is through head tracking. To detect the head, we assume that people move in a standing position, and the head is the uppermost portion of the body. To estimate the orientation of a standing person in image coordinates, we first retrieve the lowermost point of the foreground blob (which will be probably close to the ground plane). Given this position (u, v) in image coordinates, we compute the corresponding point $(x, y) = \boldsymbol{f}_w(u, v, 0)$ in world coordinates using Equation (2), making the assumption that it lies on the ground plane (i.e., $z = 0$).

Assuming a standing person, the top of the head will have world coordinates (x, y, h), where h is the height of the person (again, we assume that all persons have an average height $h_m = 1.7$m). Such point relates to a position $(u', v') = \boldsymbol{f}_i(x, y, h_m)$ in image coordinates, that can be computed directly from Equation (1). Hence we have a vector field

$$\boldsymbol{o}(u, v) = (u', v') - (u, v) = \boldsymbol{f}_i\left(\boldsymbol{f}_w(u, v, 0), h_m\right) - (u, v) \quad (3)$$

that provides the expected orientation in image coordinates for a standing person whose bottom (foot) is located at pixel (u, v).

Fig. 2 illustrates the vector field $\boldsymbol{o}(u, v)$ overlaid to a frame captured by our camera. It can be observed that $\boldsymbol{o}(u, v)$ is effectively aligned with the orientation of standing people in the image.

On a frontal/lateral camera setup, we can assume that the head of the person is at the center-top of the foreground blob of the person. In this case, to estimate the head position, a histogram of the person's foreground blob (where each histogram bin corresponds to one column of the person's blob image) can be computed. The center-top of the head estimation will be given by the peak of this histogram, and the coordinates will be the bin (column) of the peak and the corresponding peak value (row).

Although this works well in some cases, such basic procedure does not work when a camera is positioned obliquely to a scene - which is our case. Fig. 3(c) shows an example were this technique erroneously estimates the head at the person's backpack. In the case of oblique cameras the body orientation must be estimated for every pixel of the image (see Fig. 2).

Once the body orientation is computed, the head position can be estimated by a similar procedure as the one described above, but rotated by the corresponding body orientation angle. Fig. 3(d) shows an example where the use of the body orientation information generates a correct estimation of the head. If blobs with large areas are detected, they are probably related to more than one person in a close distance. In such cases, local peaks of the histogram are used to extract the top of the heads. When two persons enter the scene in a close distance, only one foreground blob may be detected. In such cases, local maxima of the oriented projection that present a minimum height (1.4 meters) and are enough distant from each other (1 meter) are retrieved as different heads.

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

Fig. 3: Top of head estimation. a) Original object; b) Background blob computed for the object; c) Top of head estimation with no orientation estimation; d) Top of head estimation with orientation estimation.

4) Tracking metric: One the head is estimated when a person enters the scene, it must be followed in the adjacent frames. There are several tracking algorithms using different strategies for characterizing and following the desired target, such as kernel-weighted histograms [39], [40], multiple fragments and histograms [28], or covariance matrices [29]. However, the desired targets in the proposed camera setup can be considerably small (particularly in the far field), and there are not enough pixels to estimate reliably the covariance matrix or the histogram. Instead, a simple template matching based on the Sum of Squared Differences (SSD) presented good results. To reduce computational time and tracking errors, the search region for the template T is reduced to a circular region computed based on the maximum displacement for each person in consecutive frames. This maximum displacement is set in meters, and then converted to image coordinates according to Equation (2).

To cope with appearance and illumination changes, the template T is updated every M_f frames (we used $M_f = 5$ for sequences acquired at 15 FPS). As a result of the tracking procedure, we have trajectories $(x_i(t), y_i(t))$, for $t = 1, ..., N_f(i)$, where $N_f(i)$ is the duration (in frames) of the i^{th} trajectory.

The proposed approach for people tracking presented good results, but still presents problems when strong shadows are aligned with the body's orientation, when groups enter the scene in nearby positions, and when persons with low-contrast with respect to the background appear. However, it should

be noticed that the main contribution of this work is not the tracking procedure itself (which can be easily replaced by other algorithms), but the analysis of tracked persons for event detection, as explained next.

B. Event Detection

1) Spatial Occupancy Maps: As described in [4], a Spatial Occupancy Map (SpOM) is an array that represents the spatial occupancy of an observed region. In fact, the SpOM is basically a 2D histogram that counts how many times each image pixel (u, v) was occupied by a person at some time during the training period.

The tracking algorithm described in this paper returns only the center of the template that represents each person at each frame t. To account for the dimension of the body, as well as inaccuracy in the tracking procedures, we estimate the histogram using an approach similar to Kernel Density Estimation (KDE) [41]. In KDE, a kernel centered at each observation is used to obtain an estimate of the Probability Density Function (PDF) that models the data. In this work, we employ a Gaussian kernel with standard deviation σ, where σ is selected based on the expected dimensions of a person. In [4], top-view cameras were employed, and the camera calibration procedure was trivial (and the assumed dimension of a person was the same at all positions). In this work, however, oblique cameras are employed, and the size of a person in image coordinates is highly dependent on the position of the person, so the KDE-like procedure is applied to world coordinates.

Let us consider N tracked people in the training stage, and let $N_f(i)$ denote the length of the i^{th} trajectory (in frames). Also, let $(u_i(t), v_i(t))$ denote the trajectory of i^{th} person in image coordinates. This trajectory is converted to world coordinates $(x_i(t), y_i(t)) = \boldsymbol{f}_w(u_i(t), v_i(t), 1.7)$, where $\boldsymbol{f}_w$ is the function defined by Equation (2), and 1.7m represents the expected height of the template position in world coordinates. Clearly, not all persons have the same height. To access the error when using 1.7m as a default height, we analyzed the error for all possible positions (u, v) in the image, and varying z in the range of plausible heights $[1.4, 2.0]$. Thus, the maximum error for each pixel (u, v) is given by:

$$E(u, v) = \max_z \|\boldsymbol{f}_w(u, v, 1.7) - \boldsymbol{f}_w(u, v, z)\|, \qquad (4)$$

where $\|\cdot\|$ denotes the Euclidean distance. For the scenario analyzed in this work, such error varies from 0.07m to 0.90m, the mean error (along all pixels) was 0.32m and the standard deviation was 0.17m.

The 2D histogram in world coordinates estimated through KDE is then given by

$$S^w_\sigma(x, y) = \sum_{i=1}^{N} \sum_{t=1}^{N_f(i)} g_\sigma \left(x - x_i(t), y - y_i(t)\right), \qquad (5)$$

where $g_\sigma(x, y)$ is a truncated discrete bidimensional Gaussian kernel, given by

$$g_\sigma(x, y) = \begin{cases} \frac{1}{c} e^{\frac{-x^2-y^2}{2\sigma^2}}, & \text{if } -2\sigma \leq x, y \leq 2\sigma \\ 0 & \text{otherwise} \end{cases}, \qquad (6)$$

and c is a normalization constant. In this work we used $\sigma =$ 1m, which is a rough estimate of the body diameter.

Positions where the SpOM $S^w_\sigma(x, y)$ is large may be considered "valid walkable regions". In fact,it is possible to build a binary map $I(x, y)$ such that $I(x, y) = 1$ in valid positions, and $I(x, y) = 0$ otherwise. This binary map is given by

$$I(x, y) = \begin{cases} 1 & \text{if } S^w_\sigma(x, y) \geq T_{\text{spom}} \\ 0 & \text{otherwise} \end{cases}, \qquad (7)$$

where T_{spom} is a threshold that informs the minimum spatial occupancy for a valid region. As proposed in [4], T_{spom} is computed adaptively from the SpOM by removing a portion r of the smallest values of $S^w_\sigma(x, y)$ (such discarded values are associated with the spread produced by the tail of the Gaussian, and a suggested value is $r = 0.4$).

Our hypothesis is that trajectories (or portions of the trajectory) that are considerably far from the valid walkable region may be considered unusual. Given a test trajectory obtained with the tracking procedure $(u(t), v(t))$, we compute the counterpart $(x(t), y(t))$ in world coordinates. We then compute the minimum distance

$$d_i(t) = D\left(x_i(t), y_i(t)\right), \qquad (8)$$

from the test trajectory to the valid region $I(x, y)$ across frames, and unusual portions of the trajectory are detected when $d_i(t) > T_{\text{dist}}$. Here, T_{dist} is the maximum allowed distance from the trajectory to the valid occupied region, and $D(x, y)$ is the Distance Transform [42] of the binary map $I(x, y)$. Although we believe that T_{dist} is context-dependent, a suggested value is $T_{\text{dist}} = 2$m.

An example of the SpOM is illustrated in Fig. 4. Fig. 4(a) shows the filmed environment along with captured trajectories, and Fig. 4(b) illustrates the SpOM. Figs. 4(c) and 4(d) show, respectively, the binary SpOM I and the Distance Map D. In Figs. 4(b)-(d), the functions $S^w_\sigma(x, y)$, $I(x, y)$ and $D(x, y)$ were mapped to image coordinates for a better visualization.

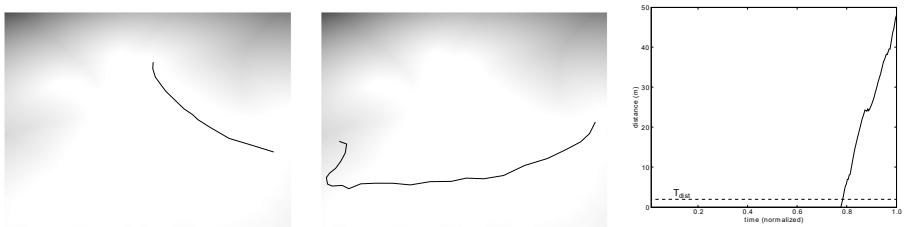
Two analysis of trajectories are illustrated in Fig. 5. Figs. 5(a) and 5(b) show two examples of trajectories overlaid to the Distance Map, while Fig. 5(c) show the distance function $d_i(t)$ for Fig. 5(b). Again, in Figs. 5(a) and (b) all functions were mapped to image coordinates for clarity. The first trajectory was considered usual at all points, since its distance from the valid region was always lower than T_{dist}, hence its function plot is not shown. On the other hand, the second trajectory was considered unusual near to the end. This happened because the subject has entered an invalid region (the grass patch at the left side of the camera image).

2) Interpersonal Relationships: The method described in the previous section captures unusual behavior only based on the spatial occupancy of a given trajectory, but does not consider the possible relationships among tracked people. In this work, we explored interpersonal relationships by analyzing the formation and classification of groups based on sociological concepts, such as *proxemics*, comfort distances, etc, as described in [5].

The term *proxemics* has been firstly proposed by Edward Hall [43] in order to describe the social use of space (in particular, personal space). Personal space is related to the

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

Fig. 4: (a) Filmed environment and tracked trajectories, (b) SpOM, (c) Binarized SpOM, (d) Distance Transform of (c).



(a) (b) (c)

Fig. 5: (a)-(b) Trajectories overlaid to the Distance Transform, (c) Distances along the trajectory shown in (b).

area with invisible boundaries surrounding an individual's body. This area states as a comfort zone during interpersonal communication, and can disappear in specific environments or situations (e.g. elevators, dense crowds). Hall proposes four main distances (see Table I) observed in American interactions. Each distance has a particular meaning, in terms of the kind of interaction that is expected to happen. Hall argues that those meanings depend on culture, and also shows how distance constrains the types of interaction that are likely to occur.

In our approach, the position of each person (in world coordinates) is used as a site to compute a Voronoi Diagram (VD). As people move, the positions of the sites are modified in time, generating a temporal evolution of Voronoi polygons (called in this work Dynamical Voronoi Diagram, or DVD). Then, the geometry of Voronoi polygons is explored to extract and quantify sociological and psychological individual characteristics, which are used to detect the possible kinds of interactions proposed by Hall [43]. As described in [5], the evaluated characteristics are:

- The personal space (PS) for an individual is defined as the area of the corresponding Voronoi polygon. In fact, such choice matches the psychological principle of personal space, since all the points in the interior of a Voronoi polygon are closer to the site that generated this polygon than to any other site.
- The Perceived Personal Space is defined as the area of the region formed by the intersection of the vision field and the corresponding Voronoi polygon. It provides an estimate of the level of comfort, because it takes into account the field of vision of the individual and his/her PS.
- The distances from any person to neighbors is computed using the VDs. In fact, the orthogonal distance from the site of a VD to its polygon edges represents half of the distance between this site and a neighboring site. It should be noticed that the VD for a set of N sites can be computed with complexity $\mathcal{O}(N \log N)$ using a divide and conquer algorithm [44], which is much cheaper than performing an exhaustive search to compute pairwise distances.

TABLE I: Hall's classification for Personal Space.

Hall's Classification	Approximate distance	Kind of interaction
Intimate distance	up to 0.5 meters	Comforting, threatening
Personal distance	0.5 to 1.25 meters	Conversation between friends
Social distance	1.25 to 3.5 meters	Impersonal business dealings
Public distance	more than 3.5 meters	Addressing a crowd

To detect group formation, we keep track of the distance from each person to his/her neighbors (such distances are provided directly by the VD) across time. If two or more people keep short distances among them in a certain period of time (let us denote this time period T_g, measured in frames, and called *grouping period*), we consider that they form a group. In practice, even a very strong group (e.g. a married couple) can be apart during some frames, when avoiding obstacles and/or other people, but still keeping the group link. To cope with this kind of situation, we consider that two individuals form a group if they keep an intimate distance for at least a fraction p of the *grouping period* T_g, where $0 \leq p \leq 1$.

Formally, let us consider two individuals $I_i(t)$ and $I_j(t)$ at frame t, and define a binary function:

$$g(i,j,t) = \begin{cases} 1 & \text{if } d(I_i(t), I_j(t)) \leq D_{\text{intimate}} \\ 0 & \text{otherwise} \end{cases}, \quad (9)$$

where $d(I_i(t), I_j(t))$ represents the distance between agents I_i and I_j at frame t, and $D_{\text{intimate}} = 0.5$ meters is the distance

for intimate relationship, as defined in Table I. Then $I_i(t)$ and $I_j(t)$ are considered a group at frame t if:

$$\sum_{k=t-(T_g-1)}^{t} g(i,j,k) \geq pT_g. \quad (10)$$

Our experimental results indicated that 5 seconds is enough time for group formation (leading to $T_g = 75$ frames in video sequences acquired at 15 frames per second), and $p = 0.8$. However, such parameters can be fine tuned for specific applications.

We also impose our grouping property to be transitive, meaning that if I_i and I_j are grouped and I_j and I_k are also grouped, then I_j and I_k must be also grouped. In this case, I_i, I_j and I_k will be in the same group. Such transitive property is important, because we can detect large groups using only pairwise comparisons.

After detecting the formation of a group, we want to characterize it as voluntary or involuntary. Group formation may occur mainly in two situations: people can form a group whether because they want to (e.g. friends) or because they are forced to, due to lack of space (e.g. exiting a crowded football stadium). It is reasonable to think that such group characterization is related to the perceived personal spaces of the individuals in the group: friends walking together in a non-crowded environment may have plenty of PPS, but they choose to stay close. On the other hand, if several people approach a door at the same time then their PPS will be small, and they will have no choice but to stay close to other people.

To characterize a group as voluntary or involuntary, we evaluate the PPSs of all individuals of the group (which is equivalent to evaluating the comfort of individuals). It is expected that, in voluntary groups, most people should have large PPSs during the grouping time T_g. However, some individuals may walk a little behind others, and consequently their PPSs would be small, even if they belong to a voluntary group. Therefore, to detect voluntary group formations, we check if at least half of the persons in the group have sufficiently large PPSs.

More specifically, let us consider a group with N persons, formed by individuals $I_1, I_2, \cdots, I_N$. Let $c_i(t)$ be a binary "comfort" function with respect to the public distance, which returns 1 if person i is comfortable (large enough PPS), and 0 otherwise. Individual $I_i(t)$ is said to be comfortable in the previous T_g frames if:

$$\sum_{k=t-(T_g-1)}^{t} c_i(t) \geq pT_g, \quad (11)$$

where p is the same parameter used in Equation (10). The group is qualified as voluntary at frame t if at least $N/2$ elements of the group satisfy Equation (11). Otherwise, the group is characterized as involuntary.

An example of dynamic group formation can be illustrated by the following: a person rapidly approaches and reaches another person. Then he/she reduces his/her speed, so that the two persons walk side by side at an intimate distance. After some time (grouping time), a voluntary group is detected. A higher-level interpretation of this grouping behavior could be simply two friends meeting, or maybe a kidnapping situation. In fact, as we shall describe in Section III-B3 that a finite automaton could be easily implemented in the proposed model to detect sequences of events that could be related to suspect behavior, such as "approach" followed by "voluntary group". Moreover, the spatial occupancy of each tracked person could be evaluated according to the method described in Section III-B1, such that we could detect if the grouping was performed in an usual or unusual portion of space.

It is important to note that the DVD can be explored to compute several other interpersonal parameters. For instance, the temporal analysis of distances between neighboring agents can be used to detect approach or leave behaviors, and the velocity vectors can be explored to detect what kind of approach/leave is happening (e.g. from the front or behind). In fact, these simpler events can be combined for the detection of more complex behavior, as explained next.

3) Query Grammar and Finite Automaton: Sections III-B1 and III-B2 described different approaches for event detection based on spatial occupancy and interpersonal relationships, respectively. However, a combination of them may provide a powerful tool for video analysis, where a variety of complex events can be detected based on concurrent or sequential combinations of simpler events. For instance, a possible theft alarm could be issued when two people approach (one from behind), group for a while, and then one of them leaves. It should be noticed other recent papers [25], [45] also explore grammars and complex activity recognition based on simpler events, but as far as we know the present approach is the first one to combine information about spatial occupancy and interpersonal relationships (in particular, grouping effects).

The core of the proposed method is the detection of simple events (e.g. the formation and classification of groups, approach or leave movements, permanence on valid or invalid regions), and to search for concurrent or sequential combinations. Generally speaking, each kind of event can be formally described through a grammar, which should have enough flexibility to allow us to depict behaviors considering the issues commented in Section III-B2. Hence, we represent each simple event as a symbol, as illustrated in Table II. It is important to note that the list of symbols shown in Table II is not exhaustive, and can be complemented by other user-defined symbols.

In order to represent more complex behaviors, these symbols can also be combined through operators: concurrency $(+)$, sequence $(-)$ and alternation $(|)$. Using these operators, it is possible to describe behaviors in a way very similar to how we would describe them in real life, and assign a semantic meaning to them. For instance, the expression $(P)-(N)$, that means an approach behavior followed by intimate grouping, could be an indicative of a kidnapping situation (or just friends meeting). We note that the construction of a composite behavior is somewhat subjective, according to one's interpretation of it. For example, theft could be described by two distinct expressions:

- approaching from behind, involuntary intimate grouping and leaving to an invalid SpOM region: $(P+A)-(N+$

TABLE II: Symbol table for the grammar.

Behavior	Symbol
Approach/leave	P/F
From the front/behind	E/A
Valid/Invalid SpOM	V/I
Valid/Invalid DT	T/D
Social/Intimate Grouping	C/N
Public/Personal Grouping	L/S
Voluntary/Involuntary Grouping	O/R

$R) - (F + I)$

- approaching from behind, then a combination of either leaving to an invalid SpOM region or leaving to an invalid DT region: $(P + A) - ((F + I)|(F + D))$

In order to look for such behaviors within a video sequence, the expression is automatically converted to a finite deterministic automaton through a semantic analyzer. In each new video frame, the behavior of each subject is inferred, including his/her own behavior in relation to other people. As the automaton is a state machine, it "reads" the current behavior of a subject and updates its state through a function depending on the current state and behavior. When the automaton reaches its final state, an alarm is triggered, causing the system to highlight the related trajectories. Each unique complex behavior corresponds to a different automaton. Moreover, we can carry out additional queries to recorded video history.

IV. Experimental Results

This section illustrates some behaviors that can be detected using the proposed approach. All experiments were performed in the same calibrated environment, and the maps used for the spatial occupancy test (both SpOM and DT) are those shown in Fig. 4.

In the first example, a query $((P + E) - O) + V$ was formulated, aiming to detect meetings of two friends. With this query, the systems looks for pairs of tracked people that approach frontally and form a intimate group, at the same time keeping in a valid region according to the SpOM. Fig. 6 illustrates some key frames of this event. Figs. 6(a) and 6(b) illustrate frames related to "approach from the front while in a valid SpOM region" $((P + E) + V)$, while the frames in Figs. 6(c) and 6(d) relate to "form an intimate group while in a valid SpOM region" $(O + V)$.

A similar query was formulated in the second experiment, aiming to detect group formation and splitting. The search query was $P - O - F$, that relates to an approach behavior (either from the front or behind), followed by voluntary grouping, and then separation. Figs. 7 and 8 illustrate frames of two video footages in which the desired behavior was detected. In Fig. 8, one subject approached from behind, while the scenario shown in Fig. 7 relates to a frontal approach. It is interesting to note that the initial query could be refined to detect the two situations separately: $(P + A) - O - F$ would retrieve only the first situation, while $(P + E) - O - F$ would retrieve only the second one.

The final example is presented in Fig. 9. In this example, the formulated query $(I|D) - (P + A) - O - F$ was fed to the system, aiming to detect possible snatch thefts. This query tries to find a person that starts in an invalid region (either according to the SpOM or the Distance Transform regions), follows another from behind, forms a voluntary group, and leaves. Fig. 9(a) shows the first frame of the event defined by the query, where a person is located in an invalid region. The frame in Fig. 9(b) illustrates the approach from behind, and the one in Fig. 9(c) the grouping. Finally, Fig. 9(d) relates to a frame in which the possible thief is running away from his victim.

It is important to note that a great variety of higher-level events can be detected using appropriate query sentences. In fact, a semantic meaning could be assigned to one ore more sentences. For instance, the sentences $(I|D) - (P + A) - O - F$ or $(P + A) - O - (F + (I|D))$ could be related to snatch thefts (the second sentence related to a escape in an unusual region according to the SpOM or Distance Transform tests).

V. Conclusions and Future Work

This paper described a new approach for event detection based on spatial occupancy and interpersonal relationships. A tracking algorithm suited for an oblique/tilted camera was introduced to compute the trajectories of people, which are used to build a Spatial Occupancy Map of the scene in a training period. In the test period, several interpersonal events are computed, and they can be combined with the spatial occupancy criterion through a search query for the detection of a wide variety of events.

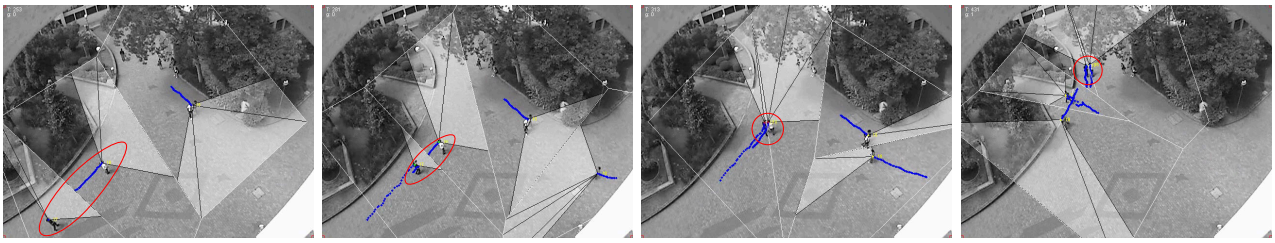
Although we did not focus on any specific application, we believe that the proposed approach may be applied to different scenarios, from surveillance to the understanding of group formation and classification. However, we believe that the main problem for using our approach in widespread environments is the limitation caused by tracking issues, particularly in crowded scenes. Despite the relative efficiency of the proposed tracking algorithm, there are still situations in which the tracking will fail, and the use of multiple cameras [46] might improve tracking results.

It is also important to note that the grouping algorithm requires the computation of distances in world coordinates, which is highly dependent on the accuracy of the tracking algorithm. Nevertheless, these errors will usually show for people far from the camera, hence not being a significant concern.

As future work, the grammar could be extended in a number of ways. For instance, at present we do not consider group attributes, such as number of people, group area, approximation and leaving time, etc. With these, we could determine, for instance, if the approach was very quick or in a rather slow way, or whether there are many people involved or not.

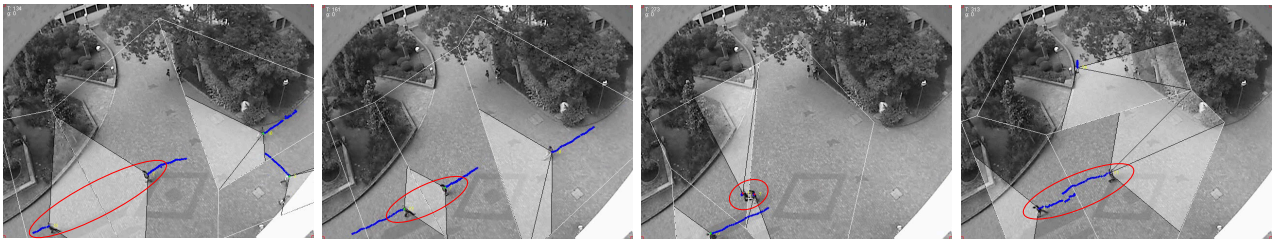
Another possibility is to also consider the body position of individuals. Similarly to other papers (e.g. [25], [47], we could analyze if an individual is standing, sitting, standing still, lying down or even fighting with other(s). This could also be fed into the grammar, in order to better portray some behaviors.

Finally, we envisage the use of additional acquisition equipment to assist in the SpOM building process. For instance, night vision cameras could greatly help during evenings, where



(a) Frame 252 (b) Frame 280 (c) Frame 312 (d) Frame 430

Fig. 6: Grouping behavior: (a)-(b) two individuals are approaching; (c) individuals are in intimate distance; (d) individuals keep moving together.

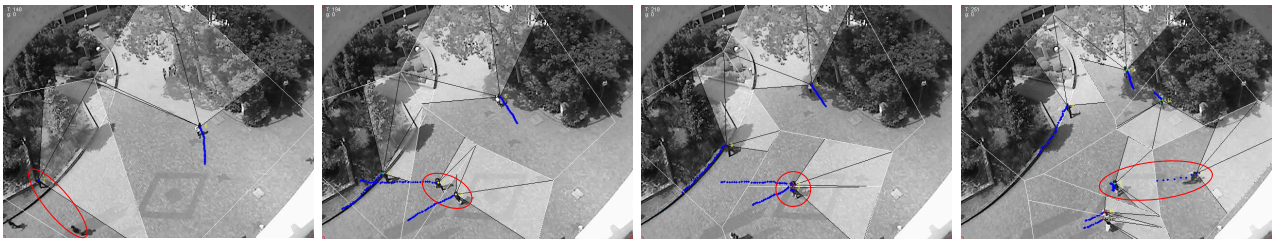


(a) Frame 133 (b) Frame 160 (c) Frame 272 (d) Frame 312

Fig. 7: Grouping behavior (frontal): (a)-(b) two individuals are approaching; (c) individuals meet and stay in intimate distance for a while; (d) individuals are moving away from each other.

(a) Frame 150 (b) Frame 170 (c) Frame 200 (d) Frame 292

Fig. 8: Grouping behavior (from behind): (a)-(b) two individuals are approaching; (c) individuals meet and stay in intimate distance for a while; (d) individuals are moving away from each other.



(a) Frame 147 (b) Frame 193 (c) Frame 217 (d) Frame 250

Fig. 9: Possible theft behavior: (a) note the individuals coming from the bottom of the frame and from an invalid SpOM region - this can indicate that the second individual adopts a suspect behavior; (b) the two individuals are approaching; (c) the individuals are in intimate distance and have stopped; (d) after a short period, the individuals start to go in different directions and one of them is running.

traditional tracking methods usually fail due to the excessive noise present in the camera image. These additional sources could enhance the system reliability.

REFERENCES

[1] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man, and Cybernetics - Part C*, vol. 34, no. 3, pp. 334–352, August 2004, doi:10.1109/TSMCC.2004.829274 .

[2] T. B. Moeslund and A. H. and. V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 1, pp. 90–126, October 2006, doi:10.1016/j.cviu.2006.08.002.

[3] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on*

Circuits and Systems for Video Technology, vol. 18, no. 11, pp. 1473–1488, 2008, doi:10.1109/TCSVT.2008.2005594.

[4] C. R. Jung, J. C. S. Jacques Jr., J. Soldera, and S. R. Musse, "Detection of unusual motion using computer vision," in *Proceedings of XIX Brazilian Symposium on Computer Graphics and Image Processing*. Manaus, Brazil: IEEE Press, September 2006, pp. 349–356, doi:10.1109/SIBGRAPI.2006.11.

[5] J. C. S. Jacques Jr., A. Braun, J. Soldera, S. R. Musse, and C. R. Jung, "Understanding people motion in video sequences using voronoi diagrams," *Pattern Analysis and Applications*, vol. 10, pp. 321–332, 2007, doi:10.1007/s10044-007-0070-1.

[6] N. Haering, , P. L. Venetianer, and A. Lipton, "The evolution of video surveillance: an overview," *Machine Vision Applications*, vol. 19, no. 5-6, pp. 279–290, 2008, doi:10.1007/s00138-008-0152-0.

[7] G. Lavee, E. Rivlin, and M. Rudzsky, "Understanding video events: a survey of methods for automatic interpretation of semantic occurrences in video," *IEEE Transactions on Systems, Man Cybernetics - Part C*, vol. 39, no. 5, pp. 489–504, 2009, doi:10.1109/TSMCC.2009.2023380 .

[8] I. Junejo, O. Javed, and M. Shah, "Multi feature path modeling for video surveillance," in *IEEE International Conference on Pattern Recognition*, 2004, pp. II: 716–719, doi:10.1109/ICPR.2004.594.

[9] M. Fuentes and A. Velastin, "Tracking-based event detection for cctv systems," *Pattern Analysis and Applications*, vol. 7, no. 4, pp. 356–364, 2004, doi:10.1007/s10044-004-0236-z.

[10] L. Fuentes and S. Velastin, "People tracking in surveillance applications," *Image and Vision Computing*, vol. 24, no. 11, pp. 1165–1171, November 2006, doi:10.1.1.108.716.

[11] H. Weiming, X. Xiao, Z. Fu, and D. Xie, "A system for learning statistical motion patterns," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1450–1464, 2006, doi:10.1109/TPAMI.2006.176.

[12] H. Weiming, X. Dan, F. Zhouyu, Z. Wenrong, and M. Steve, "Semantic-based surveillance video retrieval," *IEEE Transactions on Image Processing*, vol. 16, no. 4, pp. 1168–1181, 2007, doi:10.1.1.468.3592.

[13] C. R. Jung, L. Hennemann, and S. R. Musse, "Event detection using trajectory clustering and 4-d histograms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 1, pp. 1565–1575, 2008, doi:10.1109/TCSVT.2008.2005600.

[14] X. Wang, K. T. Ma, G.-W. Ng, and W. E. L. Grimson, "Trajectory analysis and semantic region modeling using a nonparametric bayesian model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, june 2008, pp. 1–8, doi:10.1.1.407.9889.

[15] B. T. Morris and M. M. Trivedi, "Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2287–2301, 2011, doi:10.1109/TPAMI.2011.64.

[16] W. Hu, X. Li, G. Tian, S. Maybank, and Z. Zhang, "An incremental dpmm-based method for trajectory clustering, modeling, and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 5, pp. 1051–1065, 2013, doi:10.1109/TPAMI.2012.188.

[17] H. Buxton and S. Gong, "Advanced visual surveillance using bayesian networks," in *IEEE International Conference on Computer Vision*, Cambridge, Massachusetts, June 1995, pp. 111-123, doi:10.1.1.261.3258.

[18] R. Hosie, S. Venkatesh, and G. West, "Classifying and detecting group behaviour from visual surveillance data," in *IEEE International Conference on Pattern Recognition*, 1998, pp. Vol I: 602–604, doi:10.1109/ICPR.1998.711215.

[19] N. Oliver, B. Rosario, and A. Pentland, "A bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831–843, August 2000, doi:10.1109/34.868684.

[20] Y. Du, G. Chen, W. Xu, and Y. Li, "Recognizing interaction activities using dynamic bayesian network," in *IEEE International Conference on Pattern Recognition*, vol. 1, August 2006, pp. 618–621, doi:10.1109/ICPR.2006.977.

[21] S. Gong and T. Xiang, "Recognition of group activities using dynamic probabilistic networks," in *IEEE International Conference on Computer Vision*. Washington, DC, USA: IEEE Computer Society, 2003, p. 742, doi:10.1.1.65.4779.

[22] T. Xiang and S. Gong, "Beyond tracking: Modelling activity and understanding behaviour," *International Journal of Computer Vision*, vol. 67, no. 1, pp. 21–51, 2006, doi:10.1007/s11263-006-4329-6.

[23] X. Wang, X. Ma, and W. E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 539–555, 2009, doi:10.1109/TPAMI.2008.87.

[24] X. Liu and C. Chua, "Multi-agent activity recognition using observation decomposed hidden markov models," *Image and Vision Computing*, vol. 24, no. 2, pp. 166–175, February 2006, doi:0.1016/j.imavis.2005.09.024.

[25] M. S. Ryoo and J. K. Aggarwal, "Semantic representation and recognition of continued and recursive human activities," *International Journal of Computer Vision*, vol. 82, no. 1, pp. 1–24, 2009, doi:10.1007/s11263-008-0181-1.

[26] W. Ge, R. T. Collins, and R. B. Ruback, "Vision-based analysis of small groups in pedestrian crowds," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 5, pp. 1003–1016, 2012, doi:10.1109/TPAMI.2011.176.

[27] Z. Cheng, L. Qin, Q. Huang, S. Yan, and Q. Tian, "Recognizing human group action by layered model with multiple cues," *Neurocomputing*, 2014, in press, doi:10.1016/j.neucom.2014.01.019 .

[28] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Proceedings of IEEE Computer Vision and Pattern Recognition*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 798–805, doi:0.1109/CVPR.2006.256.

[29] F. Porikli, O. Tuzel, and P. Meer, "Covariance tracking using model update based on lie algebra," in *IEEE Computer Vision and Pattern Recognition*, 2006, pp. I: 728–735.

[30] D. Ramanan, D. A. Forsyth, and A. Zisserman, "Tracking people by learning their appearance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 65–81, 2007, doi:10.1.1.81.3347, doi:10.1109/TPAMI.2007.250600.

[31] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 3457–3464, doi:10.1109/CVPR.2011.5995667.

[32] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1820 –1833, 2011, doi:10.1109/TPAMI.2010.232.

[33] W. Choi, C. Pantofaru, and S. Savarese, "A general framework for tracking multiple people from a moving camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, 2012, doi:10.1109/TPAMI.2012.248.

[34] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409 –1422, 2012, doi:10.1109/TPAMI.2011.239.

[35] J. Bins, L. L. Dihl, and C. R. Jung, "Target tracking using multiple patches and weighted vector median filters," *Journal of Mathematical Imaging and Vision*, vol. 45, no. 3, pp. 293–307, 2013, doi:10.1007/s10851-012-0354-y.

[36] G. Führ and C. R. Jung, "Combining patch matching and detection for robust pedestrian tracking in monocular calibrated cameras," *Pattern Recognition Letters*, 2014, doi:10.1016/j.patrec.2013.08.031.

[37] J. C. S. Jacques Jr., C. R. Jung, and S. R. Musse, "A background subtraction model adapted to illumination changes," in *IEEE International Conference on Image Processing*. Atlanta, GA: IEEE Press, 2006, pp. 1817–1820, doi:10.1109/ICIP.2006.312599.

[38] E. Davies, *Machine Vision: Theory, Algorithms, Practicalities, Third Edition*. Morgan Kaufmann, 2005.

[39] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, 2003, doi:10.1109/TPAMI.2003.1195991.

[40] Z. Fan, M. Yang, and Y. Wu, "Multiple collaborative kernel tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 7, pp. 1268–1273, 2007, doi:10.1.1.161.4605.

[41] J. Hwang, S. Lay, and A. Lippman, "Nonparametric multivariate density estimation: a comparative study," *IEEE Transactions on Signal Processing*, vol. 42, no. 10, pp. 2795–2810, October 1994, doi:10.1.1.51.3992.

[42] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Addison-Wesley Publishing Company, 1992.

[43] E. T. Hall, *The Silent Language*. Garden City, NY: Doubleday Company, 1959.

[44] F. Aurenhammer, "Voronoi diagrams a survey of a fundamental geometric data structure," *ACM Computing Surveys*, vol. 23, no. 3, pp. 345–405, 1991, doi:10.1145/116873.116880 .

[45] L. Lin, H. Gong, L. Li, and L. Wang, "Semantic event representation and recognition using syntactic attribute graph grammar," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 180–186, 2009, doi:10.1145/116873.116880 .

[46] S. M. Khan and M. Shah, "Tracking multiple occluding people by localizing on multiple scene planes," *IEEE Transactions on Pattern*

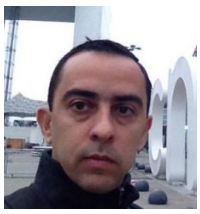
Analysis and Machine Intelligence, vol. 31, no. 3, pp. 505–519, 2009, doi:10.1.1.159.1730.

[47] F. Cupillard, A. Avanzi, F. Bremond, and M. Thonnat, "Video understanding for metro surveillance," in *Networking, Sensing and Control, 2004 IEEE International Conference on*, vol. 1, 2004, pp. 186–191, doi:10.1109/ICNSC.2004.1297432.

John Soldera received the B.Sc. degree from the Universidade de Caxias do Sul, Caxias do Sul, Brazil, in 2002, and the M.Sc. degree from the Universidade do Vale dos Sinos, São Leopoldo, Brazil, in 2007. He is currently pursuing the Ph.D. degree with the Federal University of Rio Grande do Sul, Porto Alegre, Brazil. His current research interests include biometrics, pattern recognition, and computer vision.

José Bins received the M.S. degree in Computer Science by Universidade Federal do Rio Grande do Sul (UFRGS), Brazil, in 1995, and the Ph.D. in Computer Science by Colorado State University in 2000. He held a postdoctoral position at Edinburgh University from 2004 to 2005, and is currently an Assistant Professor at Universidade Federal do Pampa (UNIPAMPA) at Alegrete in the Computer Science Department. His main research interests are on applying machine learning techniques to computer vision problems, including object tracking, behaviour recognition, and feature selection.

Marcelo Cohen is an associate lecturer at the School of Computing (PUCRS). He received his BSc from Pontifícia Universidade Católica do Rio Grande do SUL (PUCRS, 1992), his MSc from Universidade Federal do Rio Grande do Sul (UFRGS, 1996) and his PhD from the University of Leeds (2006). His current interests are game design/development and gamification, augmented reality, information and scientific visualization, and mobile applications.



Julio C. S. Jacques Junior received the M.S. degree in Applied Computer in 2006, from Universidade do Vale do Rio dos Sinos (UNISINOS) and the Ph.D. degree in Computer Sciences in 2012, from Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Brazil. He is currently a postdoctoral researcher at PUCRS. His research interests include image segmentation, shape analysis, object detection and tracking.

Soraia Raupp Musse finished her B.Sc. course in Computer Science at PUCRS, in January 1990. She earned her M.Sc. in Computer Science from UFRGS in 1994 and her Ph.D. in Computer Science from EPFL (École Polytechnique Fédérale de Lausanne) - Switzerland in 2000 supervised by Prof. Daniel Thalmann. Her research interests include crowd simulation, virtual humans animation and computer vision. She is an Associate Professor in Computer Science at PUCRS, where she supervises Posdocs researchers as well as PhD, Master and undergraduate students. She currently coordinates VHLab (Virtual Human Lab) where projects supported by private companies and the Brazilian government are developed. She has been a reviewer of important journals such as IEEE TVCG, ACM TOG and CG&A, and conferences such as ACM SIGGRAPH and Eurographics. Together with Prof. Daniel Thalmann, she is the co-author of the book "Crowd Simulation" firstly published by Springer-Verlag in 2007 and the second edition in 2012. She has authored and co-authored almost 30 peer refereed journal articles in the field of computer graphics and computer vision and more than 80 papers in conferences.

Cláudio Rosito Jung (SM11) received the B.S. and M.S. degrees in Applied Mathematics, and the Ph.D. degree in Computer Sciences, from Universidade Federal do Rio Grande do Sul (UFRGS), Brazil, in 1993,1995 and 2002, respectively. He is currently a faculty member at UFRGS in the Computer Science department. His research interests include medical imaging, multiscale image analysis, inteligent vehicles, object tracking, multimedia applications, human motion analysis, audiovisual signal processing and stereo/multiview matching.