Neural Processing Letters manuscript No.
(will be inserted by the editor)

A novel prediction method for early recognition of global human behaviour in image sequences

Jorge Azorin-Lopez · Marcelo Saval-Calvo · Andres Fuster-Guillo · Jose Garcia-Rodriguez

Received: date / Accepted: date

Abstract Human behaviour recognition has been, and still remains, a challenging problem that involves different areas of computational intelligence. The automated understanding of people activities from video sequences is an open research topic in which the computer vision and pattern recognition areas have made big efforts. In this paper, the problem is studied from a prediction point of view. We propose a novel method able to early detect behaviour using a small portion of the input, in addition to the capabilities of it to predict behaviour from new inputs. Specifically, we propose a predictive method based on a simple representation of trajectories of a person in the scene which allows a high level understanding of the global human behaviour. The representation of the trajectory is used as a descriptor of the activity of the individual. The descriptors are used as a cue of a classification stage for pattern recognition purposes. Classifiers are trained using the trajectory representation of the complete sequence. However, partial sequences are processed to evaluate the early prediction capabilities having a specific observation time of the scene. The experiments have been carried out using the three different dataset of the CAVIAR database taken into account the behaviour of an individual. Additionally, different classic classifiers have been used for experimentation in order to evaluate the robustness of the proposal. Results confirm the high accuracy of the proposal on the early recognition of people behaviours.

Keywords Human behaviour recognition · Early detection · Activity representation · Neural networks · Computer vision

1 Introduction

Many different applications including ambient assisted living, video surveillance, economization of space and urban planning need the automated processing of a

Jorge Azorin-Lopez
Computer Technology Department. University of Alicante. P.O. Box 99. E-03080. Alicante. Spain
Tel.: +34 965-90-3400
Fax: +34 965-90-9643
E-mail: jazorin@dtic.ua.es

sequence of images to analyse humans in the scene. This analysis depends on the level of scene understanding required by the specific application. In the literature, the problem has been approached at different levels from single movements such as a step or a hand displacement in the lowest level, to complex activities or behaviours in the highest that need more knowledge about the context in which the system is placed. A classification of those levels of understanding can be found in [14] where four levels are proposed: motion, action, activity and behaviour from lower to upper. Despite this classification, many works treat activities and behaviours as the same.

In this paper, we are focused in the behaviour level. Different works have been carried out to solve this problem such as those reviewed in [22] and [2]. However, many of the proposals are focused on the recognition of human activities when they are completed, but not in prediction in terms of an early detection of what an individual is going to perform in the scene. The former could be approached as a problem of classifying a sequence. Nevertheless, the latter is a problem of inferring the behaviour of a person using a subset of data of the full activity that is a relatively unexplored problem. The early prediction can be useful in many applications as for example anticipating risky situations in surveillance systems, driving assistance, avoiding lack of data when occlusions occur, etc. Many studies about prediction are more focused in actions than in complex activities. Hoai and De la Torre [8] presented a method based on Structured Output SVM for early event detection. They experiment with face expressions and human actions such as walking, running, jumping, etc.

Schindler and van Gool focused on action level handling prediction [19]. They designed a system that can predict actions from videos achieving up to 90% of correct recognition by only using short snippets of 1-7 frames instead of the whole video data and with no look-ahead.

Trajectory analysis and prediction is also a current point of interest in works as of Takano et al. [20]. They propose a system that allows humanoid robots to recognize human behaviours and predict his or her future behaviours. They concatenate sequences of motion patterns as Ngram Models and use a graph to predict future behaviours. Koppula et al. presented in [10] a system to anticipate actions using an Anticipatory Temporal Conditional Random Field (ATCRF) that models the rich spatial-temporal relations through object affordances. Modelling trajectories can predict the position target where the user is going. In [25], Ziebart et al. proposed a novel approach for predicting future pedestrian trajectories using a soft-max version of goal-based planning for robot task accomplishment with people trajectories in the environment.

Human complex activities or behaviour prediction has been studied in the last decades [4], and nowadays still remains being a topic of research. It is a more complex problem due to the number of possibilities is larger compared to a complete single action prediction. Ryoo proposed in [17] the use of a "bag-of-word" that is an integral histogram to represent human activities that allows the prediction by comparing histograms. Activity forecasting term, presented in [9], carries out behaviour prediction using semantic knowledge of the scene and optimal control theory. Their experiments are focused on trajectories prediction, but the proposal has been presented for general situations. Cao et al. presented in [5] a sparse coding usage and subsamples of the sequence to predict posterior activities for partially observed sequences. Uddin et al. proposed in [24] a Human Activity

Prediction (HPA) system which uses spanning-trees to predict and recognize activities. Daily-life activities are predicted in [15]. Recently, many researchers have focused their attention in analysing and modelling driver behaviour. In [13] different multi-modal driver signals (brake/gas, pedal pressure, vehicle velocity, etc.) are processed and then employed to detect, predict and assess driving behaviour. Other related works can be found in [21, 1].

In this paper, we propose a novel prediction method able to detect human behaviour using a portion of the trajectory of a person in the scene. The method uses the Activity Description Vector (ADV) proposed in [3] as a descriptor. In this paper, the ADV have partial information of a specific activity which can belong to different parts of the behaviour. The descriptor is used as a cue of a classification stage for pattern recognition purposes. Specifically, different classifiers are trained using the ADV associated to a specific behaviour.

The main contributions to the state of the art are that the method uses a same fixed length descriptor to characterize the different activities of a person instead of temporal series and sequential information to predict activity, avoiding the need of length normalization or time adjusting for activity prediction. Moreover, the simplicity of calculating the ADV allows its use in many different situations and scenes. Finally, the early detection method uses a pattern recognition approach, being more flexible, instead of using state or semantic models that need a predefined model in order to evaluate the behaviour.

The remainder of the paper is organized as follows. Section 2 presents the novel prediction method for early recognition of human behaviour proposed in this research. Section 3 describes the datasets and samples used in the experiments as well as the classifiers. The experimental results obtained with the method are presented in Section 4 and discussed and compared to other approaches in Section 5. Finally, conclusions about the research are presented in Section 6.

2 Prediction of global human behaviour

The predictive method to early detect human behaviour is composed by different steps. First, a sequence of images is preprocessed for different purposes, mainly for noise removal (this step is not always necessary to be performed). Enhanced images, if available, or raw sequences are used as input of the main image processing tasks: segmentation and tracking [18]. The former extracts the region of interest (ROI) of each frame. As we are interested in the global behaviour conducted by a person in the scene, the ROI is the area that corresponds to a person in the image. The latter analyses which elements of a frame correspond to the same in the next one (i.e. following a person, the ROI, along the sequence). Using the tracked region of interest, a list of positions of an individual in the scene could be calculated to represent the trajectory in the sequence. The predictive model uses only the spatial trajectory information extracted by the Activity Description Vector (see Sect. 2.1).

The main focus of this paper is the model to predict global behaviour. Hence, the previous steps, related to image processing techniques to track moving objects in video sequences, are not considered here. However, as we assume a list of positions of an individual in the scene, depending on the specific application, segmentation and tracking algorithms could be critical due to they must cope with

lighting conditions, shadows, noise, etc. Dealing with moving cameras is one related important problem. Irrespectively the segmentation techniques, our method solves it making use of the spatial trajectory information calculated by means of the Activity Description Vector. This descriptor is invariant to the point of view of the camera due to the trajectory is represented on the ground plane where people are moving.

2.1 Activity Description Vector

Activity Description Vector (ADV) is a trajectory-based feature presented in [3] for representation of trajectories based on sampling the scenario instead of sampling a trajectory itself. It takes into account simple extracted features from the trajectory that correspond to a specific region of the scenario., For the sake of completeness, the ADV is presented but we refer you to [3] in order to obtain further details about its calculation.

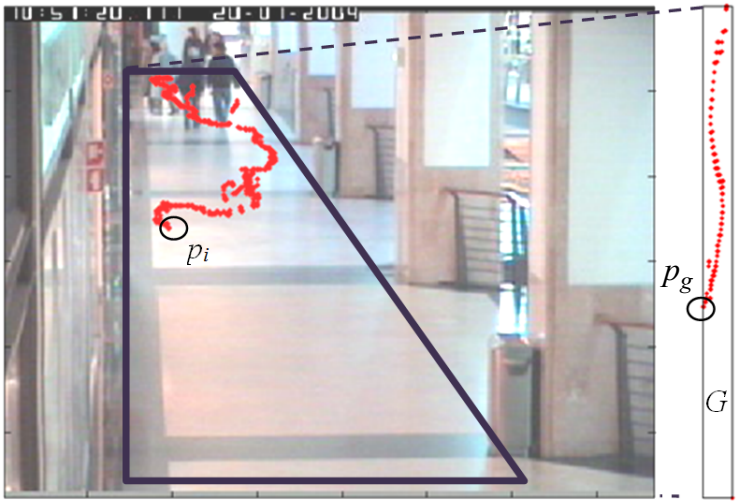


Fig. 1: Projective transformation to obtain the basic geometric model able to represent the trajectory of a person in the scenario.

The representation method takes the ground where people are moving as the basic geometric model to describe the trajectory of the individuals. We consider that data values of the scenario have to be without perspective to avoid multiple viewpoints or moving camera problems. Therefore, the space of values has to be perpendicular to the point of view of the camera. Any information contained on the image plane captured from a static camera has to be transformed to the corresponding plane that fits the ground by means of a Homography, H (1). The projective transformation allows us to consider the whole space of movements of the people in the Euclidean space (see Figure 1). Then, any point p_i on the image is transformed to a point p_g on the ground plane G.

$$p_g = H \cdot p_i \tag{1}$$

Since we are only interested in the spatial trajectory information, to obtain a simple representation to analyse the behaviour, the information needed to track the objects in the scene is the positions of an individual in the scene. They set a list of tracked points LTP on G.

$$LTP = \{p_1, p_2, p_3, \cdots, p_n\} \tag{2}$$

Typically, surveillance cameras have a frame rate of about 25 frames per second, and due to segmentation and tracking errors, the blobs that represent the analysed objects could vary in their shape. This fact produces little noisy motions that have to be avoided. Then, we propose a sampling of the LTP by taking only values of each t frames and modelling the trajectory with a spline curve, recovering a smoothed trajectory of LTP.

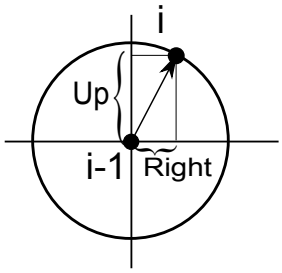


Fig. 2: Representation over axis x an y of movements Up (U) and Right (R) in a particular displacement between the point p_{i-1} and p_i.

From the smoothed tracked positions, we are able to calculate the movements of a person. Instead of calculating global positions from an origin; we consider the displacements occurred in a particular trajectory for each axis taking into account a local origin for each tracked point. Therefore, one particular movement from one tracked point to another will be calculated per each axis considering the displacement and the direction. In order to calculate it, we use four directions for each point on G: Up, U, (3), Down, D, (4), Left, L, (5) and Right, R (6). The displacement is calculated as the dot product of the displacement vector between two consecutive tracked points on LTP, p_i and p_{i-1}, and the corresponding normal vector for each axis (see Figure 2). Therefore, for a displacement of a person, movements will be:

$$U(p_i) = \begin{cases} (p_i - p_{i-1}) \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix} & if \frac{(p_i - p_{i-1}) \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix}}{\|(p_i - p_{i-1})\|} > 0 \\ 0 & otherwise \end{cases} \tag{3}$$

$$D(p_i) = \begin{cases} (p_i - p_{i-1}) \cdot \begin{bmatrix} 0 \\ -1 \end{bmatrix} & if \frac{(p_i - p_{i-1}) \cdot \begin{bmatrix} 0 \\ -1 \end{bmatrix}}{\|(p_i - p_{i-1})\|} > 0 \\ 0 & otherwise \end{cases} \tag{4}$$

$$L(p_i) = \begin{cases} (p_i - p_{i-1}) \cdot \begin{bmatrix} -1 \\ 0 \end{bmatrix} & if \frac{(p_i - p_{i-1}) \cdot \begin{bmatrix} -1 \\ 0 \end{bmatrix}}{\|(p_i - p_{i-1})\|} > 0 \\ 0 & otherwise \end{cases} \tag{5}$$

$$R(p_i) = \begin{cases} (p_i - p_{i-1}) \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} & if \frac{(p_i - p_{i-1}) \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix}}{\|(p_i - p_{i-1})\|} > 0 \\ 0 & otherwise \end{cases} \tag{6}$$

These four particular movements have information about the direction of the trajectory and the velocity of a person in a specific point on G. Additionally, we consider the frequency, F, as the number of occurrences of a person in a specific point of G. That is, the number of frames in which a person has been in a specific location. F contains information about the spatial trajectory of a person but not considering the movements themselves.

Finally, the ground plane G is spatially sampled in a matrix C of mxn cells, so that the transformed points p_g and the functions of frequency and movements of it are in one of the cells of the matrix C. Each cell will describe the activity happened in that region of the scene considering the vector of relevant values, called *Activity Description Vector* (ADV). This vector will be compound by the frequency F and the U, D, L and R movements of all points of the ground plane inside a cell:

$$ADV = \langle F, U, D, L, R \rangle \tag{7}$$

Therefore, within a particular cell, it is calculated the accumulative histograms of the movements U, D, L, R and frequency F for the points on G of the cell $C_{i,j}$ of C. Let uxv be the actual size of the scenario, mxn the number of cells in which it has been split and $p_{k,l}$ the point located in the position k and l of the G space, each ADV in a cell is described as:

$$\begin{gathered} \forall c_{i,j} \in C \wedge \forall p_{k,l} \in G / i = \left\lfloor \frac{k \text{x} m}{u} \right\rfloor \wedge j = \left\lfloor \frac{k \text{x} n}{v} \right\rfloor \\ ADV_{i,j} = \begin{pmatrix} \sum F(p_{k,l}), \sum U(p_{k,l}), \sum D(p_{k,l}) \\ \sum L(p_{k,l}), \sum R(p_{k,l}) \end{pmatrix} \end{gathered} \tag{8}$$

Hence, for a scenario space of uxv, split in mxn cells, each data in the ADV will have 5xmxn values divided in five meaningful parts with size mxn. Figure 3 shows an example of different trajectories and the ADV representation.

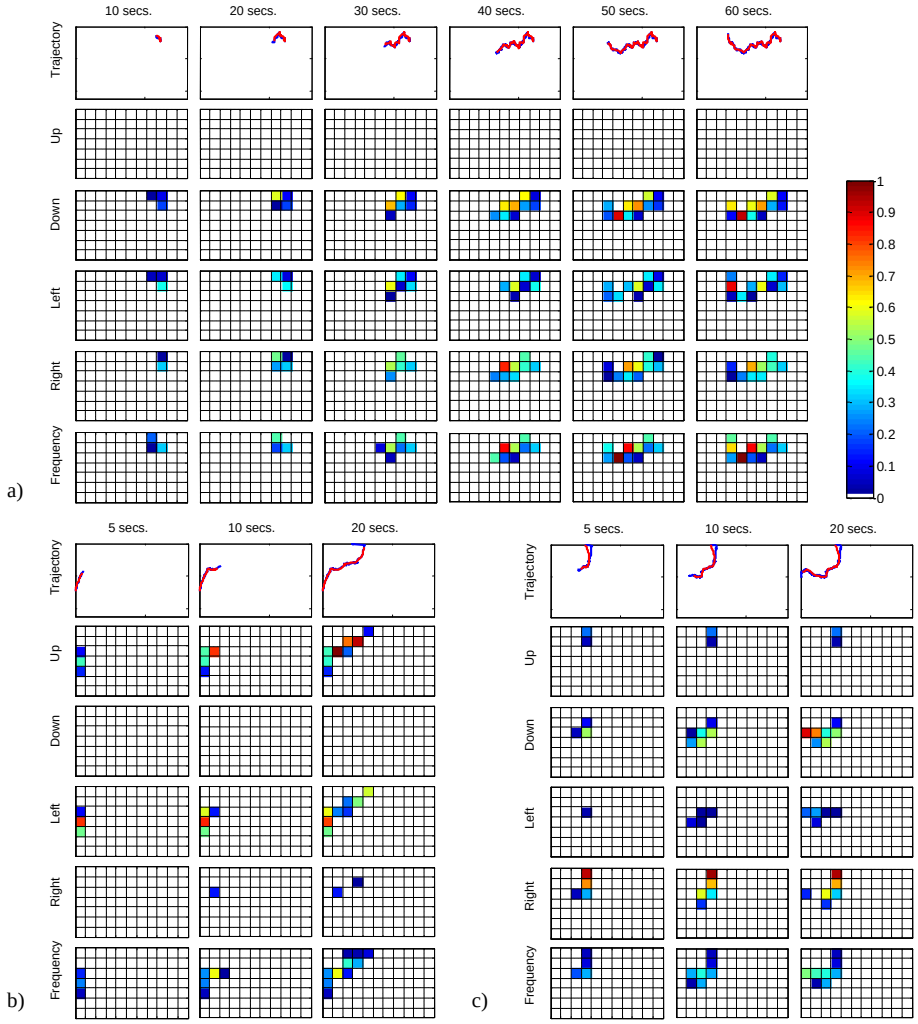


Fig. 3: ADV representation of selected samples from *Window Shopping* (a), *Shop enter* (b) and *Shop exit* (c) behaviour from *Corridor* dataset of CAVIAR for different observation times. First row shows original (blue) and smoothed (red) trajectory. The rest of rows show the Up, Down, Left, Right and Frequency that set the normalized ADV representation

2.2 Method for early prediction

The cognitive model to predict human behaviour is based on machine learning techniques. The predictive model will be able to learn from data. In this case, we are interested in learning the behaviour of a person analysing him or her in the image. The behaviour, in the highest level of understanding, is related to

complex activities and, in some cases, subtleties about knowledge to distinguish if an individual is conducting a behaviour or another very similar. For example, for the dataset used in the experiments, the difference between behaviours as browsing or window-shopping is a little nuance. Moreover, we are interested in the use of a simple representation of behaviour. Hence, the semantic gap between the input and the output could be very large.

The proposed method for early detection does not take any assumption into account about the temporal sequence of actions to model the complex behaviour. It uses a pattern recognition approach assigning a behaviour for just a trajectory instead of a predefined, semantic model of behaviour. Using predefined models for temporal sequence recognition usually requires the design of models for each behaviour derived from short-term activity patterns. For example, a window-shopping behaviour could be modelled as a sequence of states of actions in middle level of understanding as moving, browsing, moving and, finally, entering a shop [23]. Furthermore, it requires the specification of semantic understanding of low and middle level actions. Our approach is flexible in these terms as it just requires the association of high-level semantic understanding with a non-semantic input, the trajectory. This supervised learning requires that each pattern has to be labelled with a behaviour to incorporate the knowledge of a specific application. This implies some prior classification of behaviours based on the observed activity. This prior work has to be carried out by observers viewing sequences and selecting the proper labelling from a predefined semantic model and decided the characteristics of the activities that can be recognised. Supervised methods applied to the predictive model need a label for a pattern in order to translate it to a low, middle or high level behaviour. This labelling process could be relaxed using semi-supervised or even unsupervised learning techniques. For the former, just a subset of sequences is labelled. For the latter, no labelling process should be used in the learning step but it should be done for the resulted clusters.

The predictive capabilities of the proposed model are based on the generalization capabilities of the model to predict behaviour from new input samples. However, the most important predicting capability that the model provides is that it is able to detect behaviour using a portion of the trajectory of a person in the scene. The time of the subtrajectory used to predict the behaviour that a person is going to conduct in the scene will be called observation time.

The learning step uses all available samples and behaviour labels. Using the trajectory calculated from the sequence of frames by image processing techniques, the model pre-process the data (see Fig. 4). Preprocessing consists on filtering the calculated trajectory. The tracking points for individuals comprising the trajectories have usually some variations in pixels positions due to segmentation errors. In order to avoid the variations, we propose a temporal sampling and calculation of a SPLINE curve from data.

The next step in the pipeline of the model is calculating the Activity Description Vector [3]. For learning, the model calculates and stores in a database the ADV for all available trajectories using the whole trajectory including labels corresponding to each behaviour. This database is used as an input of an off-line learning process. For all available samples, a normalization is carried out in order to make the ADV independent to the observation time (i.e. independent to the trajectory length). Each ADV sample is normalized to the range (0, 1) dividing each component of the vector by the maximum value for each component in all

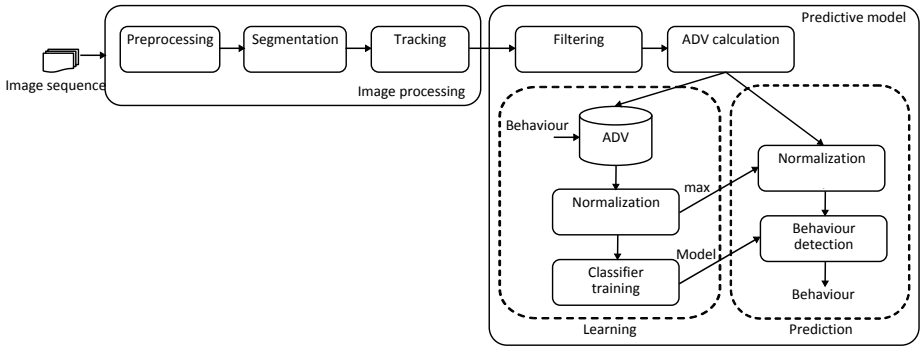
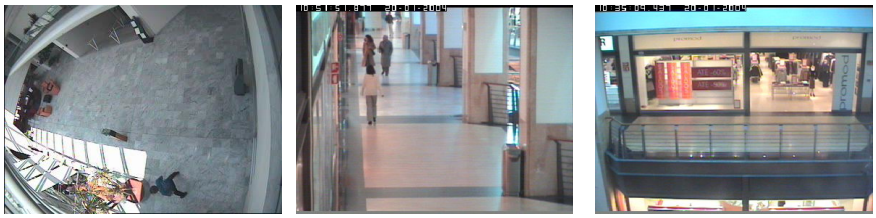


Fig. 4: Overview of the proposed predictive model

available samples. Finally, the normalized ADV is used as an input cue for the classifier training step. It is based on the pattern recognition paradigm as a problem of classification.

The predictive model step uses the same pipeline but it is able to predict the behaviour while a person is moving in the scene. Therefore, the ADV is calculated while the image sequence is processed to calculate the list of tracked points. Again, the ADV is normalized taken into account the maximum values for each component of the ADV calculated on the learning process. Finally, the classification model is used to recognize the behaviour. In this case, we are working on behaviour prediction. In consequence, as a problem of early detection or recognition, we are focused in the problem of inferring the behaviour using a subset of data of the full activity.

It is important to highlight that the proposed model does not require a subset of the trajectory to train the system. The predictive method uses the whole trajectory associated to a behaviour in the learning step. In consequence, it is not necessary to take into account subsequences of a specific behaviour to train the system making easier to learn behaviours. It has not taken into account characteristics about the length of the trajectory that must be used to correctly classify the behaviour from the activity. For the prediction step, as the system is able to distinguish the behaviour using a subset of the trajectory of a person (while he or she is moving in the scene), it is possible to classify behaviours that can be contained in a more complex behaviour. For example, walking could be an individual behaviour or a part of a window-shopping depending on the observation time of this last behaviour. In our approach, it is not considered as we do not use a sequence of actions to stablish the complex behaviour. It is just taken into account the label provided to the whole trajectory. However, depending how long the trajectory has been observed, the system could label a behaviour that needs less observation time. In this specific example, the system could predict just walking for a specific observation time until the system has enough information from the trajectory to properly predict the complex behaviour (in Sect. 3 more examples can be found). Although, we have considered these cases as a wrong prediction in the experiments, the predictive model would be having a plausible result. Anyway, the results are closely related to the knowledge provided by the observer as we stated before.



(a) Inria frame (b) Corridor view frame (c) Frontal view frame

Fig. 5: Frames from the different image datasets

3 Materials and methods

3.1 Datasets and samples

Experiments have been carried out using the CAVIAR database [6] that is available for use by the computer vision community. It contains two datasets: Inria and Shopping Centre. The first dataset was recorded in the entrance lobby of the INRIA Labs at Grenoble, France (see Fig. 5a) with image sequences of 384x288 pixels at 25 frames per second. The Shopping Centre dataset contains different clips (at the same resolution and frame rate as before) from a shopping centre in Portugal recorded from two points of view: *Corridor* view (see Fig 5b) and *Frontal view* (see Fig 5c).

The datasets were labelled manually frame-by-frame, tracking each individual in the sequences using a unique identifier. Therefore, each frame has a set of tracked individuals visible in that frame that are surrounded by a bounding box and labelled according to the situation in which the individual is involved.

Each tracked individual has a set of labels for different levels of understanding that describes the context, the situation, the movement and the role (see Table 1). The context is unique for each tracked person and involves the individual in a sequence of situations. The person has also been labelled according how much he or she is moving and the role that takes in the sequence.

Table 1: Different levels of understanding in the datasets

Dataset	Movement	Role	Situation	Context
Inria	active inactive running walking	browser fighter leaving object walker	browsing inactive moving	browsing drop down Immobile walking
Shopping centre	active inactive walking	browser walker	browsing inactive moving shop enter shop exit	browsing Immobile shop enter shop exit shop reenter walking windowshop

The goal of the experimentation is validating the proposed model to predict complex behaviour using a simple representation calculated from the trajectory of an individual person. In consequence, we only take into account the context label of the CAVIAR sequences as the high-level interpretation of the behaviour of a person in the scene. This information is subjective and depends on the observer. Additionally, we use the bounding box positions as the low-level data to describe the tracked trajectory of a person. In this case, the information is objective but noisy. There are some variation in it due to the labelling was done by humans, but we can consider it as ground truth and avoid to perform our segmentation and tracking.

The Inria dataset contains 28 clips of people walking, browsing, leaving objects, collapsing, meeting and fighting. In total, it has 26.419 frames capturing 139 individuals at 25 frames per second. As we are interested in the context information for an individual, the dataset contains four different behaviours: *Walking*, *Browsing*, *Inmobile* and *Drop down*. For each context, the average time for a person performing a specific activity is about 17 seconds (see Table 2).

Since the Shopping Centre dataset was recorded at the same time from 2 different views, the total number of people and the labelled behaviours are different. The *Corridor* dataset contains information about behaviours and trajectories performed in a long corridor with different stores. In total, 235 persons were labelled in the 26 labelled clips performing 255 different trajectories (some persons have different contexts for the sequence). The trajectories have been used as samples classified into 7 contexts: *Browsing*, *Immobile*, *Shop enter*, *Shop exit*, *Shop reenter*, *Walking* and *Windowshop* (Table 2). However, the *Frontal view* dataset contains information about a specific part of the corridor (a store) having, in consequence, less people and trajectories for the same behaviours: 144 samples.

Table 2: Samples and sequence time used in experiments

Dataset	Context	Samples	Average (secs)	Std. (secs.)	Min. (secs.)	Max. (secs.)
Inria	Walking	73	7.55	4.97	0.12	21.12
	Browsing	11	23.78	8.90	12.80	41.72
	Inmobile	51	13.84	9.64	0.56	42.20
	Drop down	4	24.65	3.12	21.08	28.60
Corridor	Shop Enter	55	13.80	10.36	0.68	58.24
	WindowShoping	18	44.77	26.62	7.44	93.40
	Shop Exit	63	16.21	13.28	0.32	48.76
	Shop Reenter	5	6.07	2.58	3.48	9.28
	Browsing	10	30.02	15.83	3.96	51.16
	Inmobile	22	22.92	22.38	0.12	79.28
	Walking	82	23.00	18.83	0.88	72.24
Frontal	Shop Enter	22	12.74	5.91	4.56	22.16
	WindowShoping	4	24.26	8.97	13.00	34.24
	Shop Exit	34	14.42	5.03	4.28	24.40
	Shop Reenter	3	23.05	8.58	15.60	32.44
	Browsing	15	30.43	29.90	5.84	112.24
	Inmobile	3	22.41	19.53	11.04	44.96
	Walking	63	7.68	3.49	0.36	17.08

The time spent for the individuals performing a specific behaviour vary notoriously in the 3 datasets having high standard deviation values for all behaviours. For example in the *Corridor*, a person takes in average about 15 seconds for short sequences as *Shop enter*, *Shop exit*, even about 6 seconds for *Shop reenter*. However, more than a half minute for long sequences as *WindowShopping* or *Browsing* can be found. For the *Frontal*, average time is similar to the previous one except on *Walking* and *WindowShoping* that are shorter. In general, the sequences in *Inria* are shorter than in Shopping Centre. The longest sample is 42 seconds for an immobile person in *Inria*, 93 seconds for an individual who is window-shopping in *Corridor* and 112 seconds for a sample of *Browsing* behaviour in *Frontal*. Short samples lasting less than a second are related to a bad labelling process; however they are taken into account to incorporate some noisy data in the process.

As we mentioned before, the bounding box positions used as ROIs and the centroids of them as the tracking points for individuals comprising the trajectories have some variations in pixels positions (and consequently to the transformed positions on the plane). In order to avoid the variations, for each sequence greater than 5 seconds, the temporal data sampling is calculated at a sampling frequency of 1 Hz (i.e. we take into account the position data each 25 frames) and the SPLINE curve is calculated from the sampled data to obtain the trajectories included in each context. For sequences less than 5 seconds, raw tracked points have been used.

For all samples, the ADV has been calculated using different grid sizes: 1x1, 3x5, 5x7 and 7x11. As we can see in the Table 2, the samples are imbalanced. For example in Frontal, just the dataset has 3 samples for *Immobile* and *Shop reenter*, whereas *Walking* has 63 samples. Thus, the Synthetic Minority Over-Sampling Technique (SMOTE) [16] has been applied to obtain the same number of trajectory samples for each context: 30 ADV samples in *Inria* and *Frontal* datasets and 60 ADV samples in *Corridor*. In consequence, for behaviours with samples greater than those values are randomly downsampled.

3.2 Classifiers

In order to evaluate the robustness of the proposed method and whether the ADV descriptor can be suitable to predict the activity being performed and up to which extent of incompleteness is still reliable, the classification step will be evaluated using general classifiers. With this experimentation we want to evaluate if this method works properly even with simply classifiers. Hence, we have selected classic classifiers: Self-Organizing Map (SOM), Supervised Self-Organizing Map (SSOM) and Neural GAS (NGAS) as three different self-organized based neural networks, the Linear Discriminant Analysis (LDA) and k-Nearest Neighbour (kNN). Moreover, a multi-classifier (MC) designed from the above classifiers has been proposed. The MC calculates from an input the most frequent class classified by the mentioned classic techniques.

Specifically, the parameters for the self-organized based neural networks are the same, having 225 neurons with a Gaussian neighbourhood function to preserve the topological properties. The SOM and the SSOM have a map grid size of 15x15 using a toroidal shape. They are trained for 50 epochs. Moreover, SOM and SSOM have an additional fine-tuning for 500 epochs. The LDA is the classic linear analysis

assuming the same covariance for each label. However, a pseudo-inverse of the covariance matrix is calculated to avoid data to be not sufficient to uniquely fit a label. Finally, the kNN method uses the 3 nearest neighbours in the classification by means of the Euclidean distance.

In order to validate the predictive model capabilities according to the time a person is observed conducting a specific behaviour, a 10-fold cross validation has been performed for each grid size. The datasets have been composed by 120 ADV samples (30 samples per 4 behaviours), 420 ADV samples (60 samples per 7 behaviour) and 210 ADV samples (30 samples per 7 behaviours) for the *Inria*, *Corridor* and *Frontal* respectively. Therefore, 90, 378 and 180 randomly selected ADV samples from the 3 datasets are used as the training samples in each iteration of the cross validation and the rest of the samples are used as the validation dataset. The ADV samples of the training dataset are calculated using the whole sequence provided in the CAVIAR dataset. Fig. 6 shows different trajectories that are used to calculate the samples for the 3 datasets of CAVIAR.

The validation dataset has been selected only from the real samples assuring that each observation has been used for validation exactly once. That is, in each iteration of the cross validation, samples artificially generated by SMOTE algorithm were not used. For each element of the validation dataset, the trajectory sample has been split into subtrajectories of specific time (observation time) and the ADV is recalculated. For example, in Fig. 3, we can see three samples corresponding to *WindowShopping*, *Shop Enter* and *Shop Exit* behaviours of the *Corridor* dataset. Samples are split into sequences from 10 up to 60 seconds for the first context and from 5 to 20 seconds for *Shop Enter* and *Shop Exit* contexts in this example. Samples shorter than observation time use whole trajectory.

4 Results

Experimental results are based on the Sensitivity (correctly classified positive samples divided by the true positive samples), Specificity (correctly classified negative samples divided by the true negative samples) and Accuracy (correctly classified samples divided by the classified samples) values of the classifiers for ADV representations of different scenario sampling to validate the predictive model capabilities according to the time a person is observed conducting a specific behaviour. Additionally, the computational time, according to training and classification time, has been calculated for the *Corridor* due to it is the largest dataset. The tests has been obtained for a Matlab implementation of the classifiers running on a Intel i7 processor with 16GB of RAM.

4.1 Inria dataset

Table 3 shows the average values of classification for each Sensitivity, Specificity and Accuracy for each grid size. In bold number it is marked the best result for each grid size, and with asterisks the best result for each of the four probabilities in general. The values presented are classified depending on the observation time (from 1 second to 70 seconds). In the two shortest observations the 3x5 grid size achieves the best results. However, within 10 seconds the largest size (7x11) is the

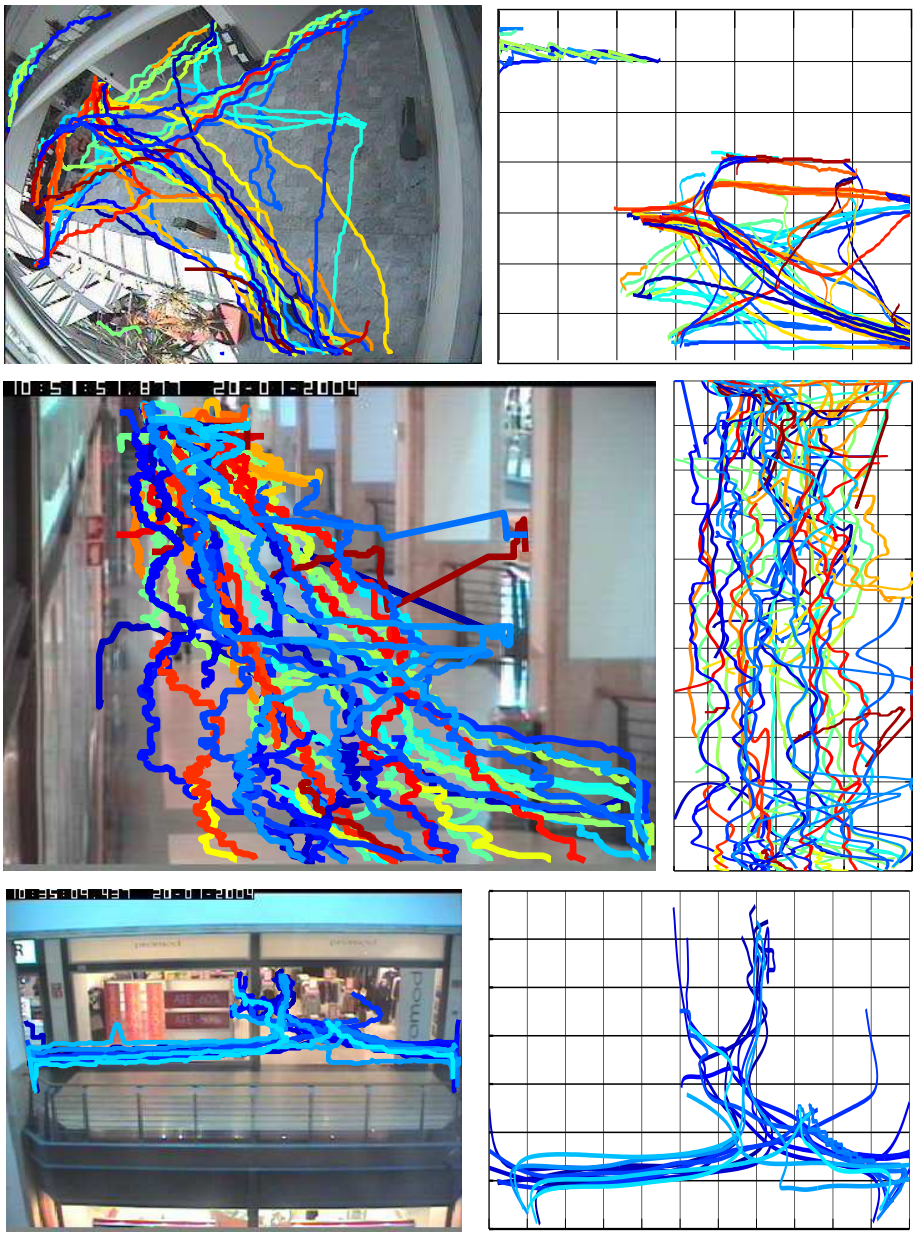


Fig. 6: Example of the *Walking* behaviour trajectories for Inria (top) and Corridor dataset (middle) in image (left) and ground plane G (right) and trajectories for the *Shop enter* behaviour of *Frontal* dataset (bottom)

best option achieving a 64% of positive classifications and over 80% of specificity and accuracy. From 20 seconds on, the best result is achieved by the 1x1 grid size

obtaining sensitivities around 75%, specificities over 90% and accuracies over 85%. In general terms, the best result in both well positive,negative classification and accuracy is achieved by 1x1 grid size in 60 seconds.

Table 3: Inria dataset. Average classification performance for different grid size and observation time

Performance	Grid	1s	5s	10s	20s	30s	40s	50s	60s	70s
Sensitivity	1x1	0.433	0.413	0.633	0.722	0.747	0.747	0.756	**0.764***	0.742
	3x5	0.467	0.520	0.633	0.689	**0.716**	0.698	0.709	0.707	0.711
	5x7	0.422	0.480	0.578	0.644	0.662	0.662	0.676	**0.678**	0.667
	7x11	0.402	0.504	0.640	0.653	**0.656**	**0.656**	0.660	0.644	**0.656**
Specificity	1x1	0.811	0.804	0.878	0.907	0.916	0.916	0.919	**0.921***	0.914
	3x5	0.822	0.840	0.878	0.896	**0.905**	0.899	0.903	0.902	0.904
	5x7	0.807	0.827	0.859	0.881	0.887	0.887	**0.892**	0.893	0.889
	7x11	0.801	0.835	0.880	0.884	0.885	0.885	**0.887**	0.881	0.885
Accuracy	1x1	0.717	0.707	0.817	0.861	0.873	0.873	0.878	**0.882***	0.871
	3x5	0.733	0.760	0.817	0.844	**0.858**	0.849	0.854	0.853	0.856
	5x7	0.711	0.740	0.789	0.822	0.831	0.831	0.838	**0.839**	0.833
	7x11	0.701	0.752	0.820	0.827	0.828	0.828	**0.830**	0.822	0.828

The study in depth of the sensitivity, specificity and accuracy according to the observation time (see Fig. 7) for 1x1 and 3x5 grid shows similar tendency to increase with the time rising. However, the values differ in sensitivity, being higher those in 1x1 grid after 25 seconds. Nevertheless, before 10 seconds the 1x1 has lower results in general terms. Particularly, LDA achieves the best result in 3x5 whereas in 1x1 has the lowest results on average. In specificity and accuracy, the classification performance follows the same behaviour than before, being better 1x1 grid size after 25 seconds. LDA has lower specificity (note that the chart represents 1-specificity, so higher values mean worse results) and accuracy in 1x1 grid size while higher in 3x5. In general terms, after 10 seconds in both grid sizes the positive classification performs over 50%, the probability of false alarm is always beneath 25% with an average $\sim 10\%$. The accuracy is over 70% on average for all classifiers in both grid sizes. The 1x1 grid has a downward tendency at the beginning because most behaviours has initially a similar trajectory, close to *walking* or *Immobile*.

Figure 8 shows the performance for the model according to each behaviour in the ROC space. The interesting case here is the *Walking* and *Immobile* in MC classifier. Initially Walking starts in the top right part of the chart, with a high rate of detection as well as high probability of false alarm. This situation occurs because initially, all behaviours are similar to walk, due to walking is included in any trajectory and hence in any behaviour except Immobile. Then, when the observation is larger, more information is added and Sensitivity achieves higher results improving the final classification.

4.2 Corridor view dataset

Table 4 shows the average results of classification performance for all classifiers in the *Corridor* dataset according to the different grid sizes (1x1 to 7x11) and the observation times (from 1 up to 70 seconds, shorter samples uses whole trajectory). Bolded values represent the best performance for each grid according to the

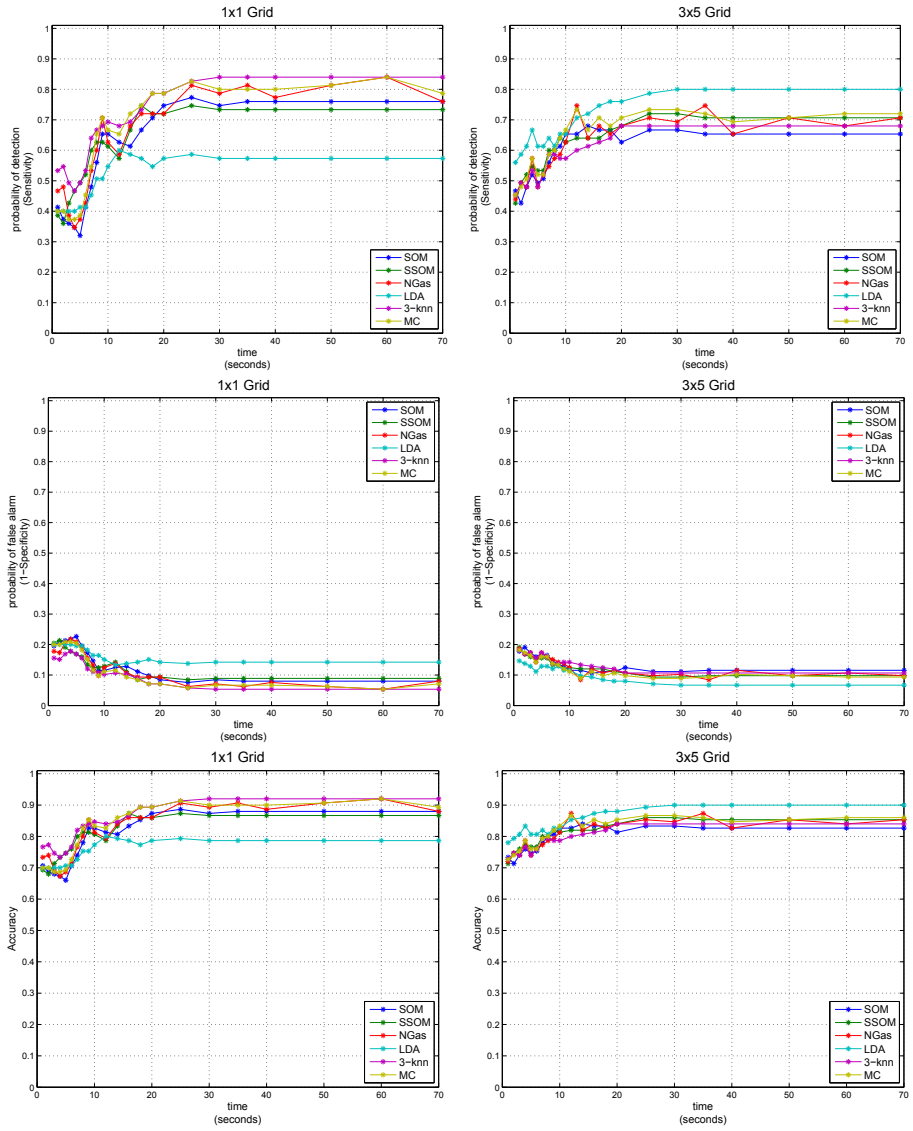


Fig. 7: Inria dataset. Probability of detection (top), false alarm (middle) and accuracy (bottom) of the method according to the observation time for 1x1 (left) and 3x5 (right) grid sizes

observation time, and those with asterisk represent the best result for each of the four probabilities in general. Best results are achieved with an observation time in 60 seconds on for a 5x7 grid. In case the system uses more data to represent the activity, grid size greater than or equal to 3x5, it requires observing a person conducting an activity less time (60 seconds) to have the highest probability to detect his or her behaviour. For observation times less than 10 seconds, the larger

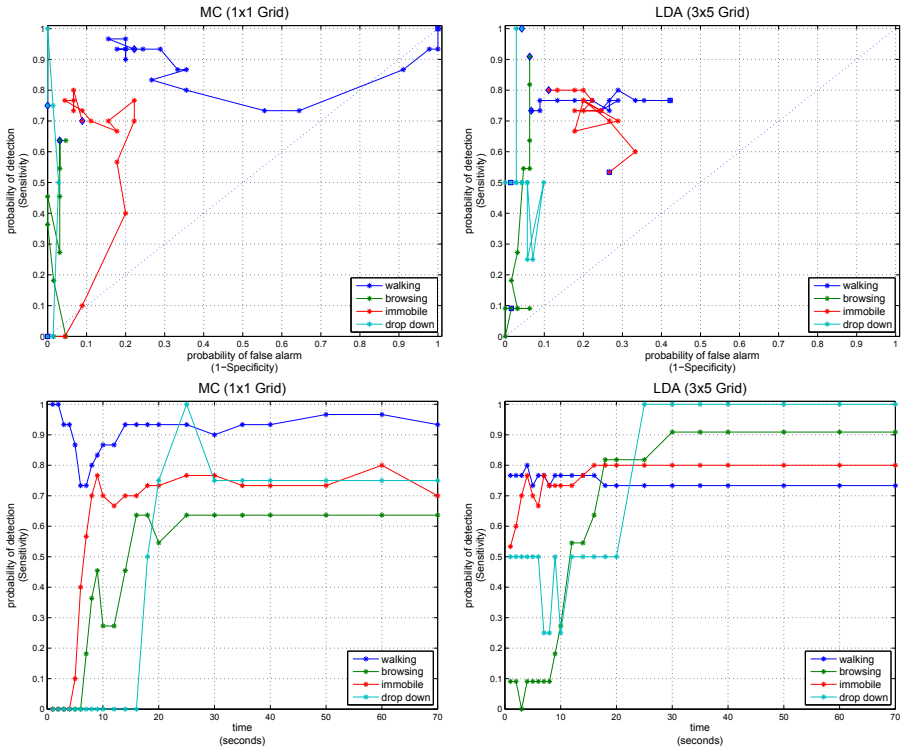


Fig. 8: Inria dataset. Performance of the predictive model in the ROC space (top) for each behaviour and the sensitivity according to the observation time (bottom) using MC (left) and LDA (right) as classifiers using a 1x1 and 3x5 grid size

grid size the better sensitivity, specificity and accuracy. However, if the observation time is greater than 10 seconds, 5x7 is the best grid size to represent the activity. Regardless of the grid size, if the observation time is greater than or equal to 10 seconds, the behaviour is correctly identified over 50% of cases keeping a very good proportion of true negative identification. From 20 seconds, the average accuracy according all classifiers is about 90%.

The study in depth of the sensitivity, specificity and accuracy according to the observation time (see Fig. 9) for 3x5 and 5x7 grid shows a similar result of the predictive capabilities of the model for both grid sizes and classifiers. The performance of the model for the *Corridor* dataset presents the best results as observation time increases. The best and worst classifier to predict behaviour depends on the grid size and observation time, being the best results, in absolute values, the LDA and the MC classifiers for an ADV calculated using a 5x7 grid and an observation time of 60 seconds. For all classifiers, except LDA, predicting a behaviour just for an observation time of 1 second has a low probability of detection ($\sim$ 15%). However, the probability of false alarm is very similar obtaining a high accuracy in the prediction ($\sim$ 80%). Finally, less than 10 seconds are enough to have a probability of 50% of proper behaviour detection, less than 10% of probability of false alarm and more than 90% of accuracy in the prediction.

Table 4: Corridor dataset. Average classification performance for different grid size and observation time

Performance	Grid	1s	5s	10s	20s	30s	40s	50s	60s	70s
Sensitivity	1x1	0.184	0.391	0.495	0.585	0.613	0.630	0.657	0.674	**0.675**
	3x5	0.181	0.464	0.577	0.687	0.738	0.762	0.765	**0.770**	**0.770**
	5x7	0.243	0.462	0.573	0.701	0.736	0.764	0.770	**0.772***	**0.772***
	7x11	0.270	0.465	0.566	0.681	0.715	0.749	0.749	**0.750**	0.749
Specificity	1x1	0.864	0.899	0.916	0.931	0.936	0.938	0.943	0.946	**0.946**
	3x5	0.864	0.911	0.930	0.948	0.956	0.960	0.961	**0.962**	**0.962**
	5x7	0.874	0.910	0.929	0.950	0.956	0.961	0.962	**0.962***	**0.962***
	7x11	0.878	0.911	0.928	0.947	0.953	0.958	0.958	**0.958**	0.958
Accuracy	1x1	0.767	0.826	0.856	0.881	0.889	0.894	0.902	0.907	**0.907**
	3x5	0.766	0.847	0.879	0.911	0.925	0.932	0.933	**0.934**	**0.934**
	5x7	0.784	0.846	0.878	0.915	0.925	0.933	0.934	**0.935***	**0.935***
	7x11	0.791	0.847	0.876	0.909	0.919	0.928	0.928	**0.929**	0.928

According to the previous results, we can conclude that if the system observes a person for more than 40 seconds, the method has a high performance. However, some samples last less than 40 seconds (see Table 2 for *Corridor* dataset). In other words, although there are samples for all behaviours (except *Shop Reenter*) which durations are larger than 40 seconds, it is necessary to study the performance of the predictive model according to the specific behaviour due to 40 seconds implies a complete behaviour process. In consequence, a study of the performance according to the observation time for each behaviour has been performed.

Figure 10 shows the performance for the model according to each behaviour in the ROC space. For all behaviours, the probability of false alarm is less than about 10% except for *Walking* that starts about 50% probability of false alarm for the LDA classifier and except for *Immobile* having a false alarm of $\sim$ 70% for 1 second of observation time if the method uses the MC classifier. In this case, the false alarm for *Walking* detection is little bit more than 10%. The probability of false alarm detecting these behaviours decreases as observation time increases. It is the most difficult to classify because all trajectories have walking component. The predictive model cannot distinguish between the generic walk and a specific walk included into another action. As we said in Sect. 2.2, we have considered this as a wrong result because we have intended to detect the complex behaviour instead of the simple action of walking. However, the person is walking in that situations as a part of the complex behaviour.

The model shows a high accuracy classifying each pattern for short observation times, being the *shop reenter* the best classified because it is the most different trajectory among the whole possible tested paths. In this case, the complex behaviour is composed by an individual who exits from a shop and enter again in less than 10 seconds. In our method, the time is not specifically considered although it can be implicitly derived from the ADV. Again, as we assume only a behaviour for sequence and we do not take into account the sequence of movements, it detects properly the *Shop exit* very quickly and takes longer to detect the actual behaviour. The shortest samples corresponding to *Shop exit* are detected using a 5x7 grid with a probability of 95% and around 1% probability of false alarm for an observation time of 3 seconds. For a 5x7 grid, after 10 seconds of observation, the system is able to detect a *Shop enter* behaviour with a probability of 50% and higher after 16 seconds, being MC the best classifier for this case. The system

classifies *Window shopping* with a probability about 16% for the LDA and 10% for the MC in 10 seconds time, raising the classification performance to a $\sim 83\%$ in 70 seconds for both classifiers.

Finally, Figure 11 shows the training and classification time according to the grid size for the ADV representation. The resulting times, average and standard deviation, has been calculated for each iteration of the cross validation test and

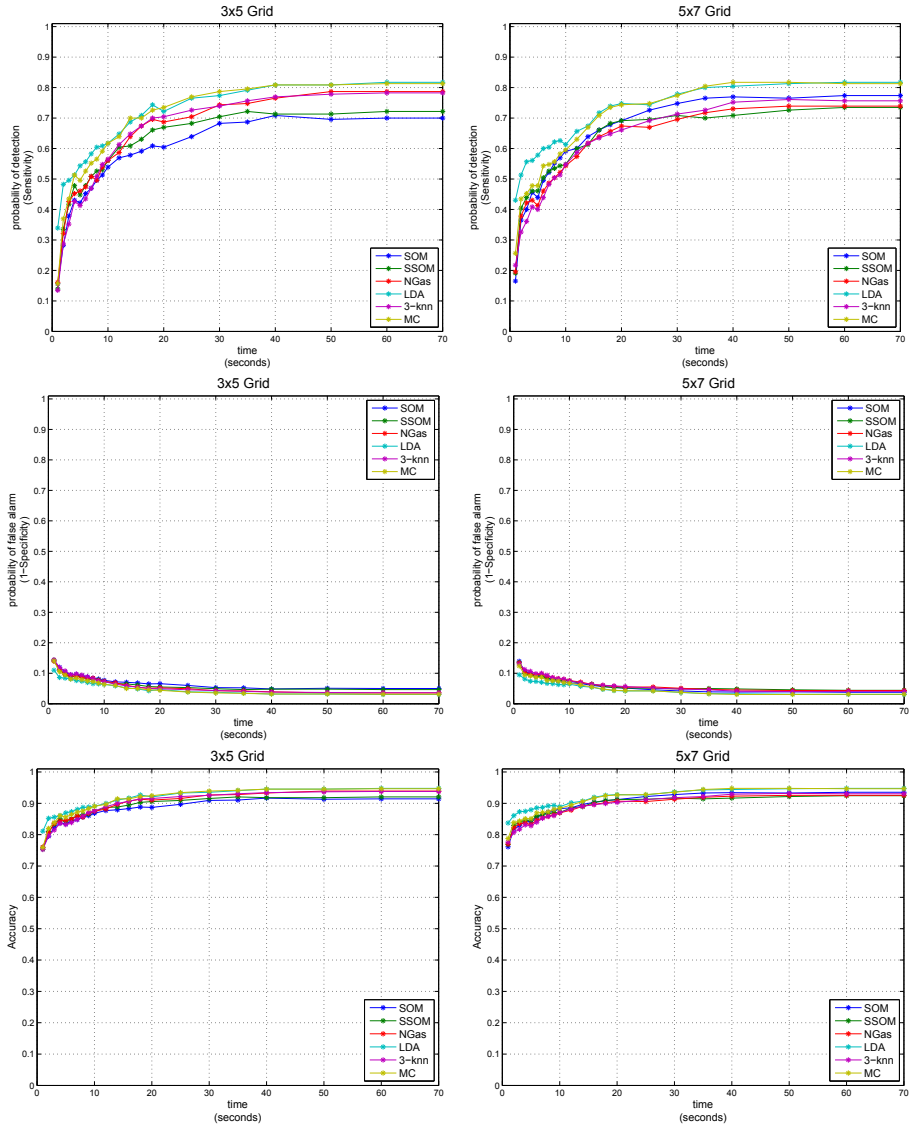


Fig. 9: Corridor dataset. Probability of detection (top), false alarm (middle) and accuracy (bottom) of the method according to the observation time for 3x5 (left) and 5x7 (right) grid sizes

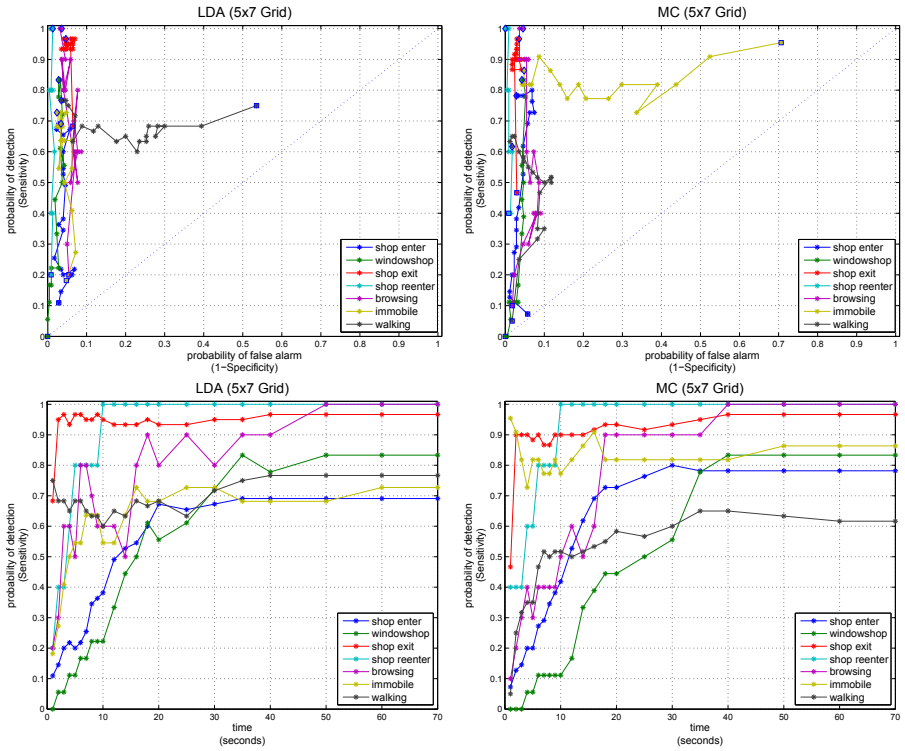


Fig. 10: Corridor dataset. Performance of the predictive model in the ROC space (top) for each behavior and the sensitivity according to the observation time (bottom) using LDA (left) and MC (right) as classifiers and a 5x7 grid size

for all considered observation times. Specifically, the training time is calculated for training the model using 378 samples for 22 different observation times (from 1 second to 60 seconds) and for 10 iterations of the cross validation. The classification time takes into acccount the remaining samples for the same number of observation times and iterations of the cross validation. The computational time of the predictive model is proportional to the grid size and the number of samples used for training. Since the ADV uses the same fixed lenght descriptor to characterize the different activities of a person, training and classification are independent on the observation time and the lenght of the trajectory. As we can see in Figure 11, SOM has the worst training and classification times, although they are insignificant according to the observation time. Within about 22 milliseconds the system is able to classify the samples for about 3.2 seconds of training period (off-line) using the largest grid size.

4.3 Frontal view dataset

Table 5 presents the average percentages of successful in classification from Sensitivity, Specificity and Accuracy terms. For all three probabilities, the best results

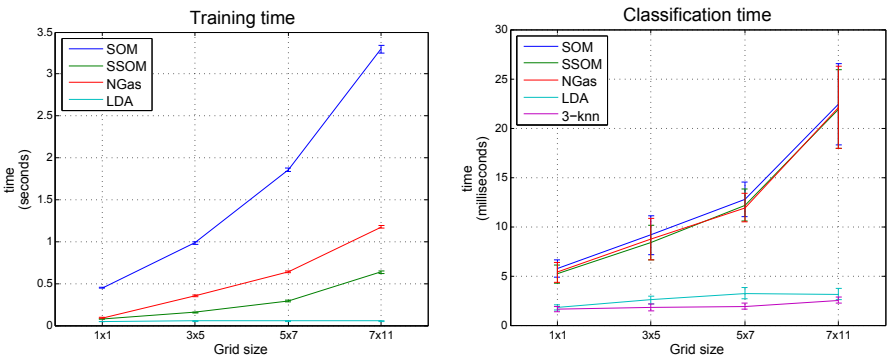


Fig. 11: Corridor dataset. Training (in seconds) and classificaction time (in milliseconds) of the predictive model for each grid size.

are achieved with 3x5 grid size except in 10 seconds that is 1x1. Moreover, using a 3x5 we achieved the best results in general, marked with an asterisk. Going in depth, within 10 second on the results on average are good, specially Specificity and Accuracy, having rates on the 90% or more. Sensitivity stars having percentages higher than 80% in 20 seconds or larger observations.

Table 5: Frontal dataset. Average classification performance for different grid size and observation time

Performance	Grid	1s	5s	10s	20s	30s	40s	50s	60s	70s
Sensitivity	1x1	0.245	0.374	0.713	0.836	0.833	0.847	0.843	**0.850**	0.846
	3x5	0.296	0.438	0.706	0.866	0.882	0.882	**0.885***	0.880	**0.885***
	5x7	0.255	0.374	0.674	0.752	**0.768**	0.762	0.757	0.759	0.765
	7x11	0.243	0.394	0.646	0.746	**0.757**	0.748	0.743	0.743	0.746
Specificity	1x1	0.874	0.896	0.952	0.973	0.972	**0.975**	0.974	**0.975**	0.974
	3x5	0.883	0.906	0.951	0.978	0.980	0.980	**0.981***	0.980	**0.981***
	5x7	0.876	0.896	0.946	0.959	**0.961**	0.960	0.960	0.960	**0.961**
	7x11	0.874	0.899	0.941	0.958	**0.960**	0.958	0.957	0.957	0.958
Accuracy	1x1	0.784	0.821	0.918	0.953	0.952	0.956	0.955	**0.957**	0.956
	3x5	0.799	0.839	0.916	0.962	0.966	0.966	**0.967***	0.966	**0.967***
	5x7	0.787	0.821	0.907	0.929	**0.934**	0.932	0.931	0.931	0.933
	7x11	0.784	0.827	0.899	0.927	**0.931**	0.928	0.927	0.927	0.927

In Figure 12 the 3x5 and 5x7 grid sizes are shown in detail for a better comprehension of the classification. In both cases the average of sensitivity is close to 90% if LDA is not taken into account. The probability of false alarm and accuracy are very similar in both sizes, but once again without LDA classifier. The general tendency is to achieve better results when the observation time is larger, achieving over 80% of sensitivity ratio after 10 seconds. The specificity is over 90% (remember that the chart shows 1-specificity) before even the 10 seconds of observation. Similarly, the accuracy achieves in 10 seconds results higher that 90%. In general terms, the classifier performances stabilize in 20 seconds achieving the best result for each one. The special case of LDA is studied next, with more concrete charts in Figure 12.

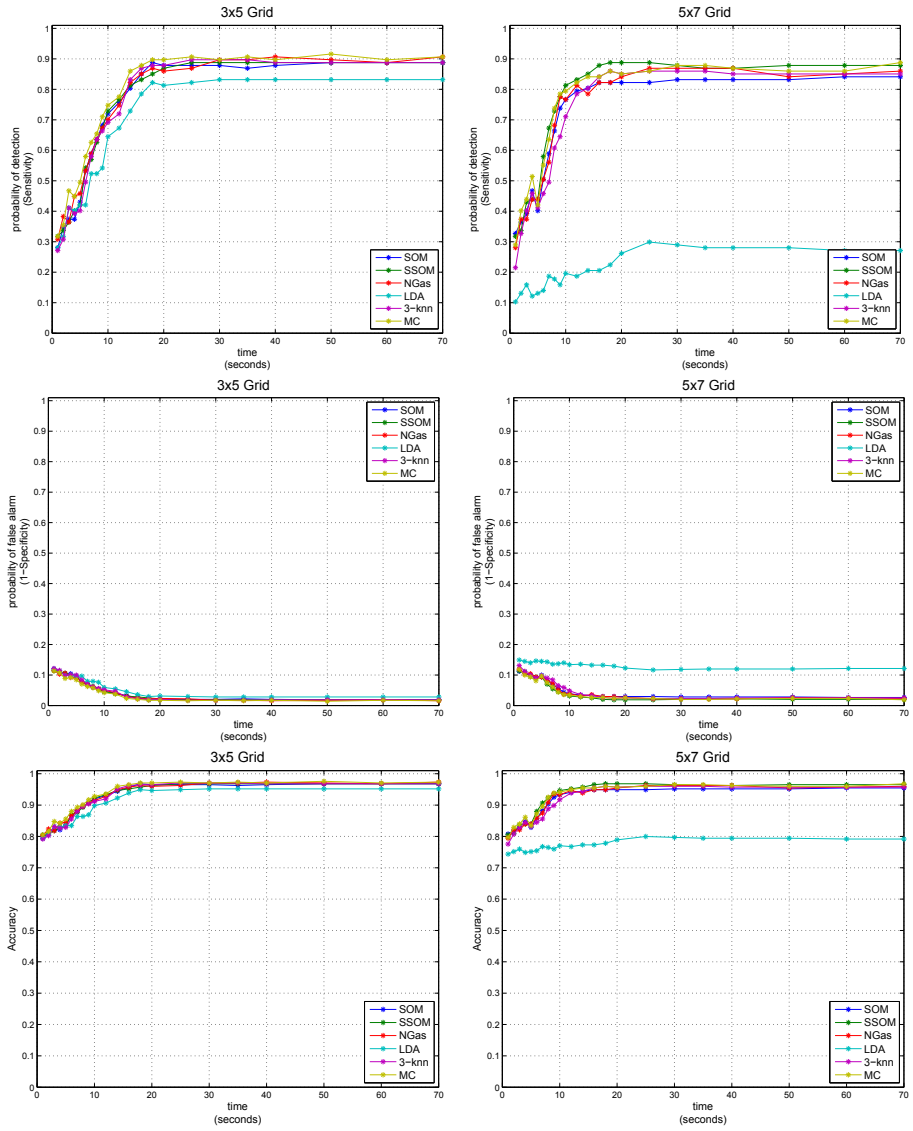


Fig. 12: Frontal dataset. Probability of detection (top), false alarm (middle) and accuracy (bottom) of the method according to the observation time for 3x5 (left) and 5x7 (right) grid sizes.

Figure 13 shows in detail LDA and MC classifiers for 5x7 grid size. The most notorious result is the LDA, that achieves very bad results. In the top left graph, the LDA in ROC space is presented. The diagonal marked with dots represents the random guess, all points above it are properly classified and under it badly predicted. In the case of LDA there is a high number of bad classification as well as random guessing. That is the reason because in the Figure 12 the LDA performed

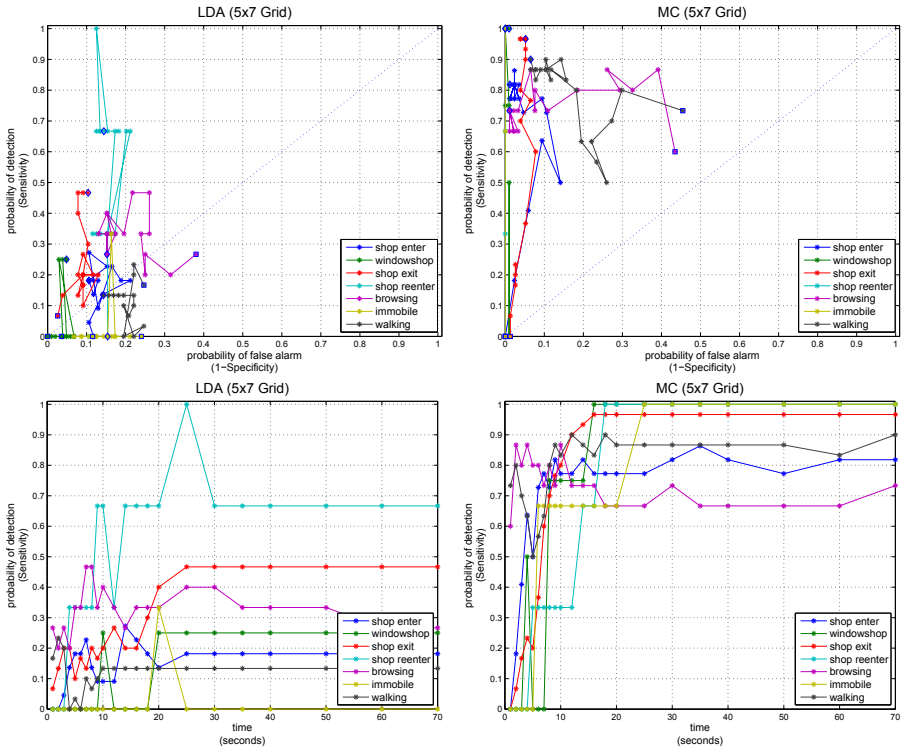


Fig. 13: Frontal dataset. Performance of the predictive model in the ROC space (top) for each behavior and the sensitivity according to the observation time (bottom) using LDA (left) and MC (right) as classifiers and a 5x7 grid size

very bad. In the left bottom chart, the LDA is presented in terms of sensitivity and time. It is easy to appreciate the low probability of detection reached even in large observations.

5 Discussion

Different conclusions can be extracted after the experimental results for the 3 different datasets. Although, the datasets have different behaviours and people involved in them, in general terms, the average results presented in Tables 3, 4, and 5 show that the Sensitivity, Specificity and Accuracy are independent of the dataset. The performance of the method is similar for all classifiers achieving best results as observation time increases. During the first 10 seconds the performance is low because the behaviours have trajectories very similar, and all can be classified as many of them. Furthermore, it tends to stabilize after 25 seconds in general. This occurs because this observation time is enough to distinguish similar behaviours. In fact, the system is able to distinguish properly the middle-level activities but we are interested in the whole behaviour performed by a person. For example, the model is able to classify a walking behaviour that could be considered a simple action

for window-shopping in the *Corridor* dataset. However, we have considered it as a wrong result because the trajectory was labelled as window-shopping. The actions carried out by a person for window-shopping needs extra time to be performed. Hence, discrimination between action, activity and behaviour is specific in different contexts and depends on the application. For the *Corridor* dataset, walking is both action and behaviour and the discrimination from the high-level of understanding is very difficult to get until the person has been observed for a long time. The best values of Sensitivity are all over 70%, Specificity higher than 90% and Accuracy over 85% for all different classifiers.

The performance of the prediction method is independent of the point of view, the trajectory of a person and the behaviour itself, but dependant of the grid size used to calculate the descriptor. The results validate the capability of the ADV descriptor to describe trajectories for early prediction purposes and the non-dependence of the classifier used for the classification step to achieve high performance in the behaviour recognition.

After studying in depth the different results (e.g. those shown in Figures 8, 10 and 13) the LDA classifier performs worse results than the rest in *Inria* and *Frontal* datasets. This situation is presented in Figure 8 left column, where the classifier has the worst result in the three studied probabilities. In both grid sizes of the Figure 13 (3x5 and 5x7) the LDA obtain bad predictive results. This results occurs in other grid sizes as well. For these datasets, the LDA is not able to calculate the pooled estimate of covariance of the samples. In consequence, it is not able to fit the multivariate normal density to each behaviour. A pseudo linear discriminant analysis had to be calculated using the Moore-Penrose pseudo inverse. Hence, classifiers based on neural networks were more robust to the method due to they were not dependant of the samples used for training.

In order to show the performance of the proposed model to predict the behaviour of a person, the MC classifier using the ADV for a 5x7 grid size has been compared to other contemporary methods in the *Corridor* dataset due to it has more people and behaviours involved. Sensitivity and specificity results of context classification have been calculated from reported success rates in [6] and [12] of comparable experiments on the same dataset. These methods are grouped as state and semantic models using predefined models and rules to evaluate behaviours.

In [6], two approaches were presented. The first, a rule-based approach, used semantic rules on both the role and movement classifications to evaluate the context from video sequences. The second, used an extension of the HMM. Specifically, to interpret the context, hidden semi-Markov model (HSMM) [23]. HSMMs extend the standard Hidden Markov model with an explicit duration model for each state [7]. Finally, in [12] Lavee et al. proposed the use of Petri Nets (PN) for recognition of event occurrences in video. The Petri Net was used to express semantic knowledge about the event domain as well as for recognizing events as they occur in a particular video sequence.

Table 6 shows results for the above three methods (Rule-based, HSMM, PN) and the proposed multi-classifier (MC) for the ADV representation using a 5x7 grid for the *Corridor* dataset. As it is shown in the table, the ADV approach achieves a significant improvement over both the Rule-based and the HSMM results for sensitivity and specificity. The predictive model outperforms the results using as a unique information the highest semantic knowledge about the behaviour (i.e. the label associated to the trajectory). Other state-of-the-art models need more

levels of semantic knowledge (low, middle and high level) to build a sequence of actions that describe the behaviour. For example, they explicitly need for a window-shopping behaviour the sequence of situations composed by moving, browsing, and then moving again associated to a specific role of a person.

Table 6: Classification performance comparison for the *Corridor* dataset

Performance	Rule-based	HSMM	PN	MC (5x7)
Sensitivity	0.57	0.6508	0.8085	0.8173
Specificity	N/A	0.9866	0.9680	0.9695

Regarding the observation time, the proposed model is able to achieve the same performance as previous works taken into account only a subset of the original sequence considering all behaviours. Concerning the probability of detection, our predictive model, using a 5x7 grid and the MC classifier, is able to achieve the 65% and the 80% of the HSMM and PN model observing a person for 11 seconds and 33 seconds respectively (see Fig. 9).

Although the performance of the methods has been calculated taking into account all behaviours, we can assume that the performance obtained by previous works, using the whole sequence, to establish the significance of the proposed model. Table 7 shows the observation time (in seconds) needed to obtain the same performance according to specific behaviours. For example, our model is able to provide the same false alarm rate than the PN model observing a person for 1 second in all behaviours, except for *Immobile*. Moreover, the method observing a person for less than 2 seconds is able to provide the same sensitivity than the HMM model for *Shop exit* and *Shop reenter* behaviours. Note that it is not possible to achieve the performance for individual behaviours marked as '-'.

Table 7: Observation time in seconds to achieve previous results for the *Corridor* dataset

Method	Performance	SHEN	WISH	SHEX	SHRE	BROW	IMMO	WALK
HSMM Sens.	65%	15	32	2	2	14	1	40
PN Sens.	81%	-	37	2	5	18	14	-
HSMM Spec.	98%	4	1	5	1	1	-	1
PN Spec.	87%	1	1	1	1	1	-	1

Acknowledgments

This work was supported in part by the University of Alicante, Valencian Government and Spanish government under grants GRE11-01, GV/2013/005 and DPI2013-40534-R.

6 Conclusions

In this paper, a predictive method to early recognize global human behaviour is proposed. The method uses the Activity Description Vector (ADV) as descriptor of the behaviour. The ADV represents trajectory of singular person in the scene by means of sampling the scenario and calculating some simple descriptors. It describes the activity happened in each region of the sampled scene. The ADV is used as a cue for different classifiers. The classifiers have as an input the ADV normalized to the range (0, 1) to be time independent. Training of the system uses the whole sequence of movements of a person and a label for the corresponding behaviour. Recognition is able to calculate the ADV of a person while he or she is performing an action in the scene. In order to validate the system, different clustering models (SOM, Supervised SOM, NGAS, LDA, kNN, MC as a combination of the others) and different grid sizes (1x1, 3x5, 5x7, 7x11) have been used. Experiments have been carried out using the CAVIAR database for 3 datasets (Inria, Corridor and Frontal). The experimental results validate the prediction capabilities of the model for any classifier and grid size. The use of classic classifiers is enough to cluster the input vectors allowing the system to correctly recognize and predict human behaviour in complex situations with great accuracy. The experimental results outperform previous works for the same dataset used in the experiments.

The proposed model is able to predict human behaviour for a short observation time by only using global information from tracking, calculated while a person is conducting the behaviour. Since the model uses only the high level semantic understanding provided in the training step to classify the behaviour, predefined models and rules to evaluate behaviours are not needed, as occurs in state and semantic models (Bayesian, HMM, Petri Nets, Grammars,...) [11]. Taking into account only the top level understanding required by the specific application, the method is able to avoid low and middle level semantic knowledge in which temporal pattern recognition methods are based to describe complex behaviours by means of a sequence of simple activities.

We are currently exploring the feasibility of the predictive model in other contexts in which other subjects (animals, biological systems, cars, etc.) describe a trajectory associated to a behaviour to analyse the generality of the model.

References

1. Angkititrakul P, Miyajima C, Takeda K (2013) Stochastic Mixture Modeling of Driving Behavior During Car Following. Journal of information and communication convergence engineering 11(2):95–102
2. Antonakaki P, Kosmopoulos D, Perantonis SJ (2009) Detecting abnormal human behaviour using multiple cameras. Signal Processing 89(9):1723–1738
3. Azorin-Lopez J, Saval-Calvo M, Fuster-Guillo A, Garcia-Rodriguez J (2013) Human Behaviour Recognition based on Trajectory Analysis using Neural Networks. In: International joint conference in neural networks, 2013
4. Beaton P, Chen Q, Meghdir H (1996) Predictive validity in stated choice studies: a before and after comparison with revealed preference. In: 1996 IEEE

International Conference on Systems, Man and Cybernetics. Information Intelligence and Systems (Cat. No.96CH35929), IEEE, vol 1, pp 205–209
5. Cao Y, Barrett D, Barbu A, Narayanaswamy S, Yu H, Michaux A, Lin Y, Dickinson S, Siskind JM, Wang S (2013) Recognize Human Activities from Partially Observed Videos. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 2658–2665
6. Fisher R, Santos-Victor J, Crowley J (2005) CAVIAR: Context Aware Vision Using Image-Based Active Recognition Project. URL http://homepages.inf.ed.ac.uk/rbf/CAVIAR/
7. Fisher R, Santos-Victor J, Crowley J (2005) CAVIAR Hidden Semi-Markov Model Behaviour Recognition. URL http://homepages.inf.ed.ac.uk/rbf/CAVIAR/hsmm.htm
8. Hoai M, De la Torre F (2012) Max-margin early event detectors. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 2863–2870
9. Kitani KM, Ziebart BD, Bagnell JA, Herbert M (2012) Activity Forecasting. In: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV, pp 201–214
10. Koppula HS, Saxena A (2013) Anticipating Human Activities using Object Affordances for Reactive Robotic Response. In: Robotics: Science and Systems (RSS)
11. Lavee G, Rivlin E, Rudzsky M (2009) Understanding video events: a survey of methods for automatic interpretation of semantic occurrences in video. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 39(5):489–504
12. Lavee G, Rudzsky M, Rivlin E, Borzin A (2010) Video event modeling and recognition in generalized stochastic petri nets. Circuits and Systems for Video Technology, IEEE Transactions on 20(1):102–118
13. Miyajima C, Angkititrakul P, Takeda K (2013) Behavior signal processing for vehicle applications. APSIPA Transactions on Signal and Information Processing 2:e2
14. Moeslund TB, Hilton A, Krüger V (2006) A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding 104(2-3):90–126
15. Mori T, Takada A, Noguchi H, Harada T, Sato T (2005) Behavior prediction based on daily-life record database in distributed sensing space. In: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, pp 1703–1709
16. N V Chawla LOH K W Bowyer, Kegelmeyer WP (2002) Smote : Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16:321–357
17. Ryoo MS (2011) Human activity prediction: Early recognition of ongoing activities from streaming videos. In: 2011 International Conference on Computer Vision, IEEE, pp 1036–1043
18. Saval-Calvo M, Azorín-López J, Fuster-Guilló A (2012) Comparative Analysis of Temporal Segmentation Methods of Video Sequences. In: Garcia-Rodriguez J, Cazorla Quevedo MA (eds) Robotic Vision, IGI Global
19. Schindler K, van Gool L (2008) Action snippets: How many frames does human action recognition require? In: 2008 IEEE Conference on Computer Vision and

Pattern Recognition, IEEE, pp 1–8
20. Takano W, Imagawa H, Nakamura Y (2011) Prediction of human behaviors in the future through symbolic inference. In: 2011 IEEE International Conference on Robotics and Automation, IEEE, pp 1970–1975
21. Tran C, Doshi A, Trivedi MM (2012) Modeling and prediction of driver behavior by foot gesture analysis. Computer Vision and Image Understanding 116(3):435–445
22. Turaga P, Chellappa R, Subrahmanian V, Udrea O (2008) Machine Recognition of Human Activities: A Survey. IEEE Transactions on Circuits and Systems for Video Technology 18(11):1473–1488
23. Tweed D, Fisher R, Bins J, List T (2005) Efficient hidden semi-markov model inference for structured video sequences. In: Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on, pp 247–254
24. Uddin MZ, Byun KM, Cho MH, Lee SY, Khang G, Kim TS (2011) A Spanning Tree-Based Human Activity Prediction System Using Life Logs from Depth Silhouette-Based Human Activity Recognition
25. Ziebart BD, Ratliff N, Gallagher G, Mertz C, Peterson K, Bagnell JA, Hebert M, Dey AK, Srinivasa S (2009) Planning-based prediction for pedestrians. In: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, pp 3931–3936