2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

# A Predictive Model for Recognizing Human Behaviour based on Trajectory Representation

Jorge Azorín-López, Marcelo Saval-Calvo, Andrés Fuster-Guilló, Antonio Oliver-Albert

*Abstract*— **The automatic understanding of the behaviour conducted by humans in scenarios using images as input of the system is a very important and challenging problem involving different areas of computational intelligence. In this paper human activity recognition is studied from a prediction point of view. We propose a model that, in addition to the capabilities of it to predict behaviour from new inputs, it is able to detect behaviour using a portion of the input. Specifically, we propose a prediction activity method based on the Activity Description Vector (ADV) to early detect the behaviour performed by a person in a scene. ADV is used to extract features that are normalized to be the cue of behaviour classifiers. We use complete sequences for training and partial sequences to evaluate the prediction capabilities having a specific observation time of the scene. CAVIAR dataset and different classic classifiers have been used for experimentation in order to evaluate the proposal obtaining great accuracy on the early recognition.**

## I. Introduction

HUMAN behaviour recognition in video sequences is an important research topic in the field of computer vision. Video surveillance, ambient intelligence, economization of space, urban planning and ambient assisted living are examples of applications in which more and more an automated behavioural analysis is needed. Different levels of understanding can be found in literature to analyze the behaviour, from single movements such as a step or a hand displacement in the lowest level, to complex activities or behaviours in the highest. A classification of those levels can be found in [1] where four levels are proposed: motion, action, activity and behaviour from lower to upper. Despite this classification, many works treat activities and behaviour as the same.

In this paper we are focused in behaviour level. Different approaches have been proposed for this purpose such as those reviewed in [2] and [3]. However, many of the proposals are focused on complete human activities recognition, but not in prediction in terms of an early detection of what an individual is going to perform in the scene.

Behaviour recognition can be seen such as a problem of classification. Nevertheless, behaviour activity prediction means inferring the behaviour using a subset of data of the full activity. Prediction can be useful in many applications as for example anticipating risking 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 [4] 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 [5]. They designed a system that can predict actions from videos achieving up to 90% of correct recognitions 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 such as the work of Takano et al. [6]. 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 [7] a system to anticipate action using an anticipatory temporal conditional random field (ATCRF) that models the rich spatial-temporal relations through objects affordances. Modelling the trajectories can predict the target where to the user is going. In [8], 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 [9], and nowadays still remains being a topic of research. It is a more complex problem due to the number of possibilities is larger comparing to a single action prediction. Ryoo proposed in [10] 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 [11], 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 [12] a sparse coding usage and subsamples of the sequence to predict posterior activities for partially observed sequences. Uddin et al. proposed in [13] a Human Activity Prediction (HPA) system which uses spanning-trees to predict and recognize activities. Daily-life activities are predicted in [14]. Recently, many researchers have focused their attention in analyzing and modelling driver behaviour. In [15] different multi-modal driver signals (brake/gas, pedal pressure, vehicle velocity, etc.) are processed and then employed to detect, predict and asses driving behaviour. Other related works can be found in [16][17].

In this paper, we propose a novel prediction method able to

J. Azorin-Lopez, M. Saval-Calvo, A. Fuster-Guilló and Antonio Oliver-Albert are with the University of Alicante, E-03080, Alicante, SPAIN (e-mail: {jazorin, msaval, fuster,aoliver}@dtic.ua.es).

This work was supported in part by the University of Alicante and Valencian Government under Grants GRE11-01 and GV/2013/005.

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

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 [18] as a descriptor. In this study, the ADV will have partial information of a specific activity which can belong to different parts of it. The main contributions to the state of the art methods are that we do not use temporal and sequential information to predict activity, avoiding the need of normalization or time adjusting for activity prediction. Moreover, the simplicity of calculating the ADV allows its use in many different situations and scenes. Therefore, one goal is to know if this descriptor can be suitable to predict the activity being performed and up to which extent of incompleteness is still reliable. This prediction method will be evaluated using general classifiers. With this experimentation we want to evaluate if this features works properly even with simply classifiers.

The remainder of the paper is organized as follows. Section II presents the ADV representation model proposed in this research that contains the proper information that allows classic classifiers to recognize human behaviour. Experiments are discussed and compared to other approaches in Section III. Finally, conclusions about the research are presented in Section IV.

## II. Prediction of Human Behaviour

The predictive model is composed by different steps. A sequence of images is preprocessed for different purposes,

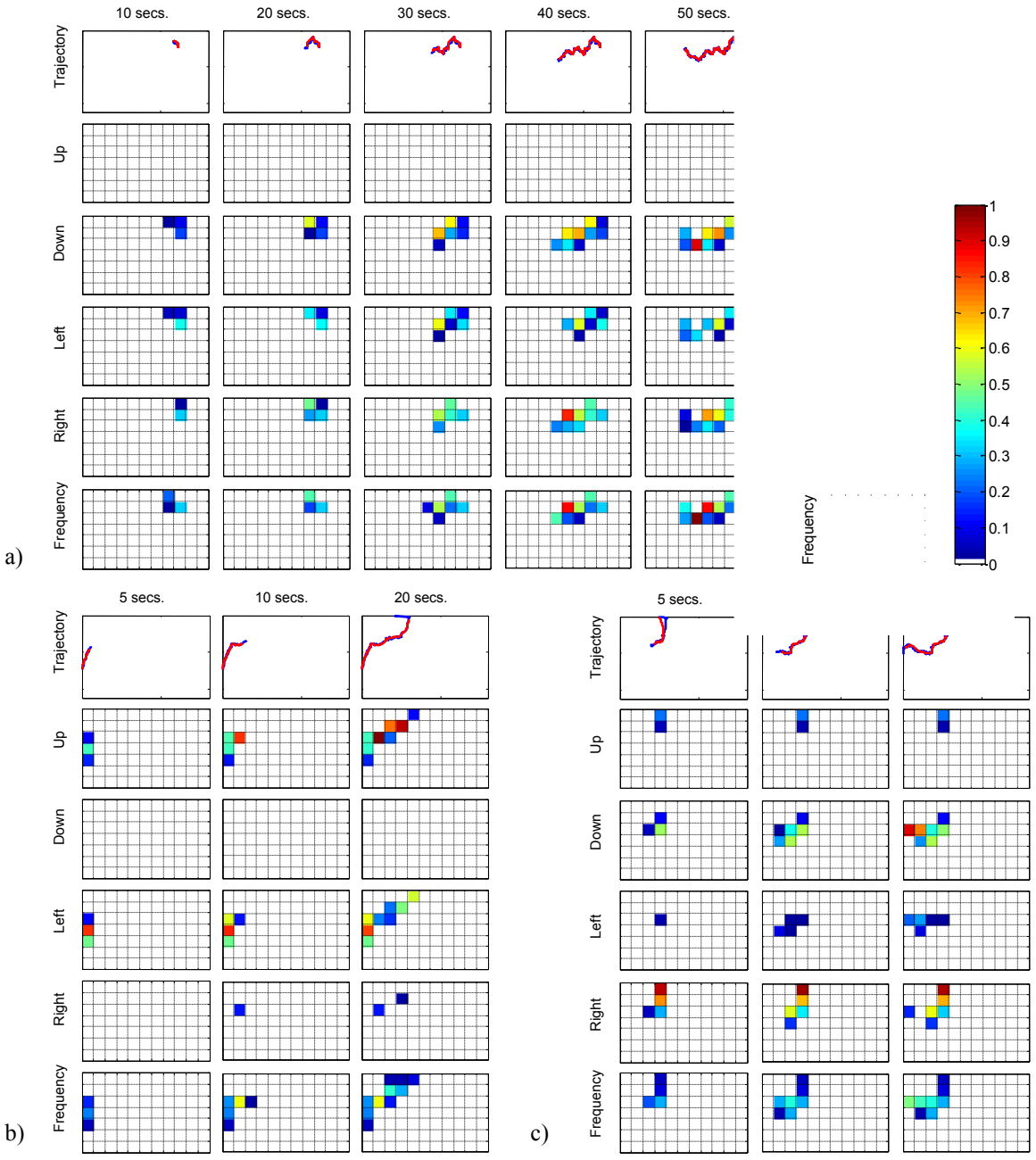


Fig. 1. Samples of *Window Shopping* (a), *Shop enter* (b) and *Shop exit* (c) behavior from CAVIAR dataset for different observation times. First row shows original (blue) and smoothed (red) trajectory. The rest of rows show the Up, Down, Left, Right and Frequency that set the normalized ADV representation.

mainly for noise removal methods. This step is not always necessary to be performed. The enhanced image, if available or the raw sequence is used as input of the main image processing tasks: segmentation and tracking [19]. The former extracts the region of interest (ROI) of each frame. As we are interested in the behaviour conducted by a person, 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, that is following the 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: the Activity Description Vector.

### A. Activity Description Vector

Activity Description Vector (ADV) is a trajectory-based feature presented in [18] for representation of trajectory data (see Fig. 1). For the sake of completeness, a brief summary of the ADV is presented but we refer you to [18] in order to obtain further details about its calculation.

ADV uses the number of occurrences of a person in a specific point of the scenario and the local movements that performs in it. This method divides the scenario, *G*, in regions as a grid, *C*, to discretize the environment. Each cell of the grid has information of the movements performed in the region that lies in it including up (*U*), down (*D*), left (*L*), right (*R*) and frequency (*F*) data. The four former values are extracted from the single displacement between two consecutive points. If we focus on the *U* movement, Equation 1 explains how this value is extracted:

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

where $p_i$ and $p_{i-1}$ are two consecutive locations of the trajectory in *G* and knowing that *U* is assumed to be a displacement in the positive vertical *y* axis. This formula is similar for the other three displacements. Frequency, *F*, on the other hand, is estimated as the number of occurrences of a person that is in a specific point. Finally, ADV is calculated within a particular cell as 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 u x v the actual size of the scenario, m x n the cells 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:

$$\forall c_{i,j}\in C\wedge\forall p_{k,l}\in G/\, i=\left\lfloor\frac{k\times m}{u}\right\rfloor\wedge j=\left\lfloor\frac{k\times n}{v}\right\rfloor$$
$$ADV_{i,j}=\begin{pmatrix}\sum F(p_{k,l}),U(p_{k,l}),\sum D(p_{k,l}),\\ \sum L(p_{k,l}),\sum R(p_{k,l})\end{pmatrix} \quad (2)$$

With this feature the trajectory is described dividing the scene in regions and compressing the data in cumulative values. It is interesting to highlight that Activity Description Vector integrates the trajectory information without length and sequential constraints, what makes it ideal for predictive purposes.

### B. Model for prediction

The cognitive model to predict human behaviour is based on machine learning area. The predictive model will be able to learn from data. In this case, we are interested on learning the behaviour of a person analyzing 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 distinguish if an individual is conducting a behaviour or another very similar. For example, for the dataset used in the experiments, the difference for behaviours as browsing or window-shopping is a little nuance. Moreover, we are interested on the use of a simple representation of behaviour. Hence, the distance of understanding between the input and the output could be very large.

The predictive capabilities of the proposed model are based on, of course, 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

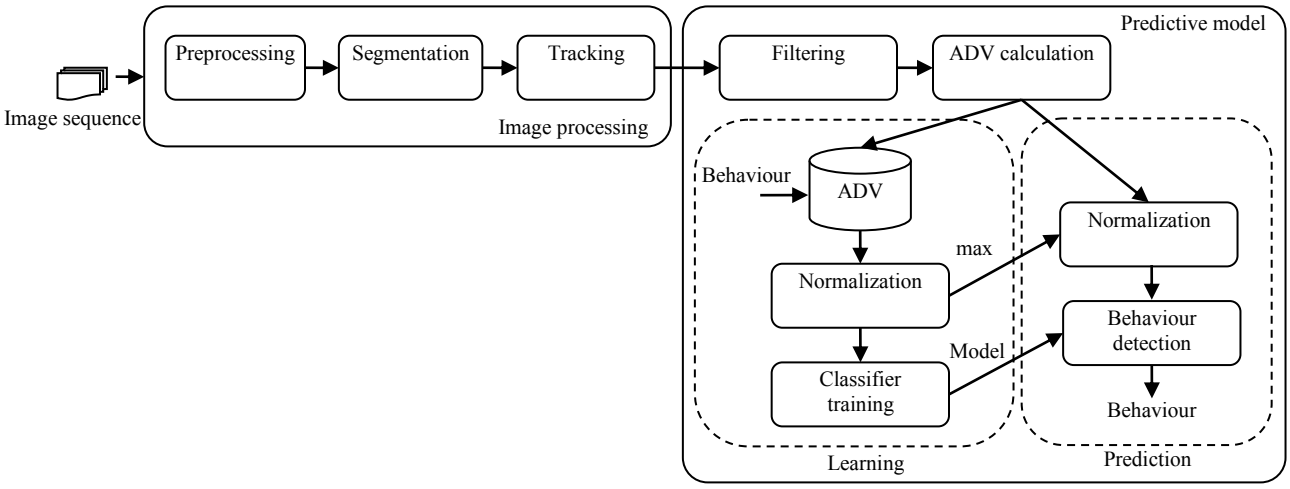


Fig. 2. Overview of the proposed predictive model

trajectory of a person in the scene. The time of the subtrajectory used to predict the behaviour a person is going to conduct in the scene will be called observation time.

The learning step uses all available samples. Using the trajectory calculated from the image by image processing techniques, the model pre-process the data (see Fig. 2). 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 [18]. For learning, the model calculates and stores the ADV for all available samples in a database including labels corresponding to each behaviour. This database is used as an input of an offline 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 available samples. Finally, the normalized ADV is used as an input cue for the classifier training.

The predictive model 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 as 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.

## III. Experiments

### A. Experimental setup

Experiments have been carried out using the CAVIAR database [20]. Specifically, validation of the predictive model to recognize human behaviours makes use of the 26 clips from the Shopping Centre in Portugal recorded from frontal view of the scenario ("view of the corridor" in the dataset). This set of sequences contains 1500 frames on average of 384x288 pixels, capturing 235 individuals at 25 frames per second. Each sequence was labelled frame-by-frame by hand and each individual is tracked using a unique identifier in the sequence. 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 have a set of labels for different levels of understanding that describes the context, the situations, the movement and the role. The context (*shop enter*, *windowshop, shop exit, shop reenter, browsing, immobile and walking*) is unique for each tracked person and involve the person in a sequence of situations (*browsing, inactive, moving, shop enter, and shop exit*). The individual also has been labelled according how much he or she is moving (*inactive, active* and *walking*) and the role that takes in the sequence (*browser* and *walker*).

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 is some variation in it due to the labelling was done by humans. The 235 persons in the 26 clips labelled from the Shopping Centre perform 255 different trajectories (some persons have different contexts for the sequence). The trajectories have been used as samples classified into the 7 contexts (Table I).

TABLE I
SAMPLES AND SEQUENCE TIME USED IN EXPERIMENTS

| Context | Samples | Average time (secs) | Min time (secs) | Max time (secs) |
|---|---|---|---|---|
| Shop Enter | 55 | 13.8 | 0.68 | 58.24 |
| WindowShoping | 18 | 44.77 | 7.44 | 93.4 |
| Shop Exit | 63 | 16.21 | 0.32 | 48.76 |
| Shop Reenter | 5 | 6.07 | 3.48 | 9.28 |
| Browsing | 10 | 30.02 | 3.96 | 51.16 |
| Inmobile | 22 | 22.92 | 0.12 | 79.28 |
| Walking | 82 | 23.00 | 0.88 | 72.24 |

The time spent for the individuals performing a specific behaviour vary notoriously. A person takes in average around 15 seconds for short sequences as 'Shop Enter', 'Shop Exit', even around 6 seconds for 'Shop Reenter'. However, more than a half minute for long sequences as 'WindowShopping' or 'Browsing' can be found. The longest sample is 93 seconds and corresponds to an individual who is windowshopping. 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 I, the samples are imbalanced. Thus, the Synthetic Minority Over-Sampling Technique (SMOTE) [21] has been applied to obtain the same number of ADV samples for each context: 60 ADV samples. In consequence, for the '*Walking'* and '*Shop Exit'* contexts, samples are undersampled randomly.

TABLE II
AVERAGE CLASSIFICATION PERFORMANCE FOR DIFFERENT GRID SIZE AND OBSERVATION TIME

| Performance | Grid | 1s | 5s | 10s | 20s | 30s | 40s | 50s | 60s | 70s |
|---|---|---|---|---|---|---|---|---|---|---|
| Sensitivity | 1x1 | 0.1841 | 0.3913 | 0.4949 | 0.5848 | 0.6130 | 0.6304 | 0.6572 | 0.6739 | **0.6746** |
| | 3x5 | 0.1812 | 0.4638 | 0.5768 | 0.6870 | 0.7384 | 0.7623 | 0.7652 | **0.7703** | **0.7703** |
| | 5x7 | 0.2428 | 0.4616 | 0.5732 | 0.7014 | 0.7362 | 0.7638 | 0.7703 | **0.7725*** | **0.7725*** |
| | 7x11 | 0.2696 | 0.4652 | 0.5659 | 0.6812 | 0.7152 | 0.7486 | 0.7493 | **0.7500** | 0.7486 |
| Specificity | 1x1 | 0.8640 | 0.8986 | 0.9158 | 0.9308 | 0.9355 | 0.9384 | 0.9429 | 0.9457 | **0.9458** |
| | 3x5 | 0.8635 | 0.9106 | 0.9295 | 0.9478 | 0.9564 | 0.9604 | 0.9609 | **0.9617** | **0.9617** |
| | 5x7 | 0.8738 | 0.9103 | 0.9289 | 0.9502 | 0.9560 | 0.9606 | 0.9617 | **0.9621*** | **0.9621*** |
| | 7x11 | 0.8783 | 0.9109 | 0.9277 | 0.9469 | 0.9525 | 0.9581 | 0.9582 | **0.9583** | 0.9581 |
| Accuracy | 1x1 | 0.7669 | 0.8261 | 0.8557 | 0.8814 | 0.8894 | 0.8944 | 0.9021 | 0.9068 | **0.9070** |
| | 3x5 | 0.7660 | 0.8468 | 0.8791 | 0.9106 | 0.9253 | 0.9321 | 0.9329 | **0.9344** | **0.9344** |
| | 5x7 | 0.7836 | 0.8462 | 0.8781 | 0.9147 | 0.9246 | 0.9325 | 0.9344 | **0.9350*** | **0.9350*** |
| | 7x11 | 0.7913 | 0.8472 | 0.8760 | 0.9089 | 0.9186 | 0.9282 | 0.9284 | **0.9286** | 0.9282 |

Regarding the classifier step, experiments have been carried out using classic classifiers: Self-Organizing Map (SOM), Supervised Self-Organizing Map (SSOM), Neural GAS (NGAS), Linear Discriminant Analysis (LDA) and k-Nearest Neighbour (kNN). Moreover, a multiclassifier (MC) designed from the above classifiers has been designed. The MC calculates from an input the most frequent class classified by the mentioned classic techniques.

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 dataset has been composed by 420 ADV samples (60 ADV samples per behaviour). Therefore, 378 randomly selected ADV samples are used as the training dataset 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.

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 Figure 1, we can see three samples corresponding to 'WindowShopping', 'Shop Enter' and 'Shop Exit' behaviours. 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.

### *B. Results and discussion*

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.

Table II shows the average results of classification accuracy for all classifiers according to the different grid size (1x1 to 7x11) and the observation time (from 1 up to 70 seconds, shorter samples uses whole trajectory). Bolded values represent the best performance for each grid according to the observation time. Best results are achieved with an observation time of 70 seconds for a 1x1 grid. In case the

TABLE III
CLASSIFICATION PERFORMANCE FOR 5X7 GRID SIZE AND DIFFERENT OBSERVATION TIME

| Performance | Classifier | 1s | 5s | 10s | 12s | 20s | 30s | 40s | 50s | 60s | 70s |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sensitivity | SOM | 0.165 | 0.439 | 0.591 | 0.600 | 0.691 | 0.748 | 0.770 | 0.765 | 0.774 | 0.774 |
| | SSOM | 0.191 | 0.461 | 0.548 | 0.600 | 0.691 | 0.709 | 0.709 | 0.726 | 0.735 | 0.735 |
| | NGAS | 0.196 | 0.413 | 0.543 | 0.574 | 0.674 | 0.696 | 0.730 | 0.739 | 0.739 | 0.739 |
| | LDA | **0.430** | **0.578** | **0.613** | **0.657** | **0.748** | **0.778** | 0.804 | 0.813 | **0.817** | **0.817** |
| | KNN | 0.217 | 0.400 | 0.548 | 0.587 | 0.661 | 0.713 | 0.752 | 0.761 | 0.757 | 0.757 |
| | MC | 0.257 | 0.478 | 0.596 | 0.630 | 0.743 | 0.774 | **0.817** | **0.817** | 0.813 | 0.813 |
| Specificity | SOM | 0,861 | 0,907 | 0,932 | 0.933 | 0,949 | 0,958 | 0,962 | 0,961 | 0,962 | 0,962 |
| | SSOM | 0,865 | 0,910 | 0,925 | 0.933 | 0,949 | 0,951 | 0,951 | 0,954 | 0,956 | 0,956 |
| | NGAS | 0,866 | 0,902 | 0,924 | 0.929 | 0,946 | 0,949 | 0,955 | 0,957 | 0,957 | 0,957 |
| | LDA | **0,905** | **0,930** | **0,936** | **0.943** | **0,958** | **0,963** | 0,967 | 0,969 | **0,970** | **0,970** |
| | KNN | 0,870 | 0,900 | 0,925 | 0.931 | 0,943 | 0,952 | 0,959 | 0,960 | 0,959 | 0,959 |
| | MC | 0,876 | 0,913 | 0,933 | 0.938 | 0,957 | 0,962 | **0,970*** | **0,970** | 0,969 | 0,969 |
| Accuracy | SOM | 0,761 | 0,840 | 0,883 | 0.886 | 0,912 | 0,928 | 0,934 | 0,933 | 0,935 | 0,935 |
| | SSOM | 0,769 | 0,846 | 0,871 | 0.886 | 0,912 | 0,917 | 0,917 | 0,922 | 0,924 | 0,924 |
| | NGAS | 0,770 | 0,832 | 0,870 | 0.878 | 0,907 | 0,913 | 0,923 | 0,925 | 0,925 | 0,925 |
| | LDA | **0,837** | **0,880** | **0,889** | **0.902** | **0,928** | **0,937** | 0,944 | 0,947 | **0,948** | **0,948** |
| | KNN | 0,776 | 0,829 | 0,871 | 0.882 | 0,903 | 0,918 | 0,929 | 0,932 | 0,930 | 0,930 |
| | MC | 0,788 | 0,851 | 0,884 | 0.894 | 0,927 | 0,935 | **0,948*** | **0,948** | 0,947 | 0,947 |

system uses more data to represent the activity, grid size greater than or equal to 3x5, the system 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 more 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. Irrespectively the grid size, if the observation time is greater than or equal to 10 seconds, the behaviour is correctly identified as such over 50% of cases keeping a very good proportion of negatives which are correctly identified as such. From 20 seconds, the average accuracy according all classifiers is around 90%.

The study in depth of the sensitivity according to the observation time (see Figure 3) shows a similar result of the predictive capabilities of the model for each grid size and classifier. The performance of the sensitivity curve is similar for all classifiers getting 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 result, in absolute values, the LDA classifier for an ADV calculated using a 7x11 grid and an observation time of 60 seconds (83.04%). The worst prediction is got using the LDA and SSOM classifiers for an ADV calculated using a 1x1 grid and an observation time of 1 second (10.86%). Just 12 seconds for a 1x1 grid and 10 seconds for the other grid sizes are necessary to have a probability of 50% of proper behaviour detection.

In general, best results for all classifiers are obtained using a 5x7 grid. Table III show the sensitivity, specificity and accuracy for a 5x7 grid. Although performance results are very close for all classifiers, best results are obtained for the LDA and the MC classifier. The predictive model is able to detect behaviour with an accuracy of 90% for an observation time of 12 seconds. The best prediction for all performance values are 81.7% of sensitivity, 97% of specificity and 94.8% of accuracy. These are very good results obtained for MC and LDA classifiers from an observation time of 40 seconds for the former and 60 seconds for the latter.

According to the previous results, we can conclude that best results are obtained for an observation time of 40 seconds. However, some samples last less than 40 seconds (see Table I). 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 4 shows the performance for the model according to each behaviour in the ROC space. For all behaviours, the

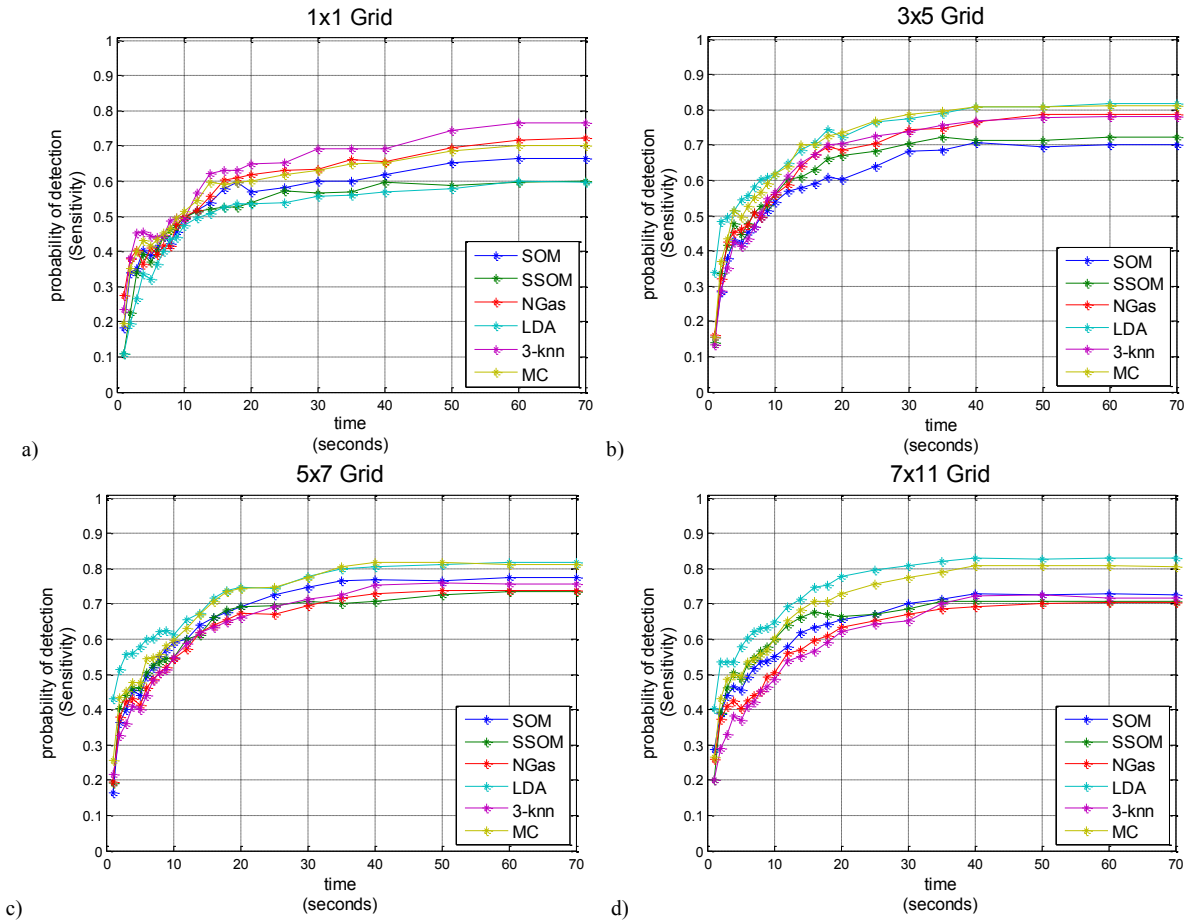


Fig. 3. Sensitivity curves according to the observation time for 1x1 (a), 3x5 (b), 5x7 (c) and 7x11 (d) grid sizes.

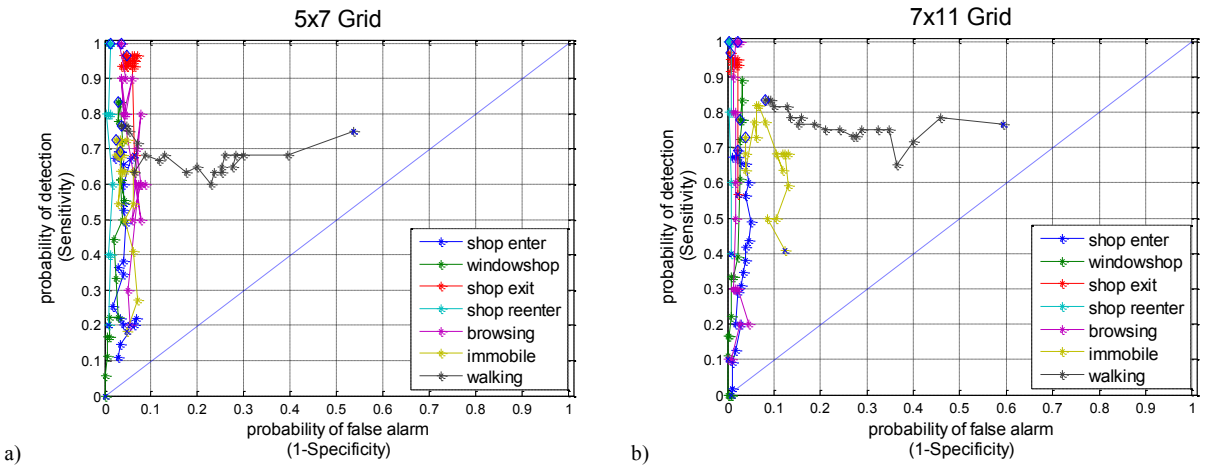


Fig. 4. Performance of the predictive model in the ROC space for each behavior using LDA as classifier and a 5x7 (a) and 7x11 (b) grid sizes.

probability of false alarm is less than about 5% except for 'walking' that starts about 60% probability of false alarm. The probability of detection of this behaviour slightly decreases as observation time increases. The shortest samples corresponding to 'shop exit' are detected using a 7x11 grid with a probability of 95% and around 0.9% probability of false alarm for an observation time of 3 seconds. For a 7x11 grid, after 10 seconds of observation, the system is able to detect 'shop enter' with a probability higher than 40% reaching a 50% and higher after 16 seconds. 'Window shopping' classifies with a probability about 16% in the first 10 seconds and raises the proper classification to a 77% in 70 seconds. For 'shop exit', 91% of classification rate has been achieved, 80% for 'shop reenter' which obtains 100% in 12 seconds. 'Browsing' has 30% percentage of Sensitivity, 63% for 'inmobile' and 75% for 'walking'. The probability of false alarms are respectively about 4%, 1%, 1.6%, 1%, 2.4%, 12.1% and 24.3% for the first 10 seconds.

TABLE IV
CLASSIFICATION PERFORMANCE COMPARISON

| | Rule-based | HSMM | PN | **LDA (7x11)** |
|---|---|---|---|---|
| Sensitivity | 0.57 | 0.6508 | 0.8085 | **0.8304** |
| Specificity | N/A | 0.9866 | 0.9680 | **0.9717** |

Model shows a high accuracy in classifying for 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. On the contrary, 'walking' could be the most difficult to classify because all trajectories have walking component. The predictive model cannot distinguish between the generic walk and a specific walk for another action.

In order to show the performance of the proposed model to predict the behaviour of a person, the LDA classifier using the ADV for a 7x11 grid size has been compared to other contemporary methods. Sensitivity and specificity results of context classification have been calculated from reported success rates in [20] and [22] 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 [20], 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 [24]. Finally, in [22] 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 IV shows results for the above three methods (Rule-based, HSMM, PN) and the proposed multiclassifier (MC) for the ADV representation using a 5x7 grid. As 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 without having semantic knowledge about behaviour.

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. Regarding the probability of detection, our predictive model is able to achieve the 65% and the 80% of the HSMM and PN model observing a person for 12 seconds and 30 seconds respectively. According to the specific behaviour, table V shows the observation time in seconds needed to obtain the same performance of previous works.

TABLE V
OBSERVATION TIME TO ACHIEVE PREVIOUS RESULTS

| | SHEN | WISH | SHEX | SHRE | BROW | INMO | WALK |
|---|---|---|---|---|---|---|---|
| 65% | 20 | 30 | 2 | **2** | **14** | **6** | **1** |
| 81% | - | - | 2 | **5** | **16** | **14** | **25** |
| 98% | 40 | - | 5 | **1** | **1** | - | - |
| 87% | 20 | 1 | 1 | **1** | **1** | - | - |

## IV. CONCLUSIONS

In this paper a predictive model to recognize human behaviour based on the Activity Description Vector (ADV) is proposed. 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. 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. 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. Best results are got using the LDA as classifier and a 7x11 grid for ADV outperforming 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 and data from tracking calculated while a person is conducting the behaviour. Predefined models and rules to evaluate behaviours are not needed in this method, as occurs in state and semantic models (Bayesian, HMM, Petri Nets, Grammars,…) [25]. We are currently exploring the feasibility of the predictive model in other contexts of human behaviour to analyse the generality of the model.

## REFERENCES


[1] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Underst.*, vol. 104, no. 2–3, pp. 90–126, Nov. 2006.

[2] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine Recognition of Human Activities: A Survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.

[3] P. Antonakaki, D. Kosmopoulos, and S. J. Perantonis, "Detecting abnormal human behaviour using multiple cameras," *Signal Processing*, vol. 89, no. 9, pp. 1723–1738, Sep. 2009.

[4] M. Hoai and F. De la Torre, "Max-margin early event detectors," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2863–2870.

[5] K. Schindler and L. van Gool, "Action snippets: How many frames does human action recognition require?," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[6] W. Takano, H. Imagawa, and Y. Nakamura, "Prediction of human behaviors in the future through symbolic inference," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 1970–1975.

[7] H. S. Koppula and A. Saxena, "Anticipating Human Activities using Object Affordances for Reactive Robotic Response," in *Robotics: Science and Systems (RSS)*, 2013.

[8] B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. Srinivasa, "Planning-based prediction for pedestrians," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 3931–3936.

[9] P. Beaton, Q. Chen, and H. Meghdir, "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)*, 1996, vol. 1, pp. 205–209.

[10] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *2011 International Conference on Computer Vision*, 2011, pp. 1036–1043.

[11] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Herbert, "Activity Forecasting," in *12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV*, 2012, pp. 201–214.

[12] Y. Cao, D. Barrett, A. Barbu, S. Narayanaswamy, H. Yu, A. Michaux, Y. Lin, S. Dickinson, J. M. Siskind, and S. Wang, "Recognize Human Activities from Partially Observed Videos," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2658–2665.

[13] M. Z. Uddin, K. M. Byun, M. H. Cho, S. Y. Lee, G. Khang, and T.-S. Kim, "A Spanning Tree-Based Human Activity Prediction System Using Life Logs from Depth Silhouette-Based Human Activity Recognition," in *Computer Analysis of Images and Patterns, 14th International Conference*, 2011, pp. 302–309.

[14] T. Mori, A. Takada, H. Noguchi, T. Harada, and T. Sato, "Behavior prediction based on daily-life record database in distributed sensing space," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005, pp. 1703–1709.

[15] C. Miyajima, P. Angkititrakul, and K. Takeda, "Behavior signal processing for vehicle applications," *APSIPA Trans. Signal Inf. Process.*, vol. 2, p. e2, Mar. 2013.

[16] C. Tran, A. Doshi, and M. M. Trivedi, "Modeling and prediction of driver behavior by foot gesture analysis," *Comput. Vis. Image Underst.*, vol. 116, no. 3, pp. 435–445, Mar. 2012.

[17] P. Angkititrakul, C. Miyajima, and K. Takeda, "Stochastic Mixture Modeling of Driving Behavior During Car Following," *J. Inf. Commun. Converg. Eng.*, vol. 11, no. 2, pp. 95–102, Jun. 2013.

[18] J. Azorin-Lopez, M. Saval-Calvo, A. Fuster-Guillo, and J. Garcia-Rodriguez, "Human Behaviour Recognition based on Trajectory Analysis using Neural Networks," in *International joint conference in neural networks, 2013*, 2013.

[19] M. Saval-Calvo, J. Azorín-López, and A. Fuster-Guilló, "Comparative Analysis of Temporal Segmentation Methods of Video Sequences," in in *Robotic Vision*, J. Garcia-Rodriguez and M. A. Cazorla Quevedo, Eds. IGI Global, 2012.

[20] R. Fisher, J. Santos-Victor, and J. Crowley, "CAVIAR: Context Aware Vision Using Image-Based Active Recognition Project." [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/. [Accessed: 01-May-2013].

[21] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE : Synthetic Minority Over-sampling Technique," vol. 16, pp. 321–357, 2002.

[22] G. Lavee, M. Rudzsky, E. Rivlin, and A. Borzin, "Video Event Modeling and Recognition in Generalized Stochastic Petri Nets," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 1, pp. 102–118, 2010.

[23] D. Tweed and R. Fisher, "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*, 2005, pp. 247 – 254.

[24] R. Fisher, J. Santos-Victor, and J. Crowley, "CAVIAR Hidden Semi-Markov Model Behaviour Recognition." [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/hsmm.htm. [Accessed: 01-May-2013].

[25] G. Lavee, E. Rivlin, and M. Rudzsky, "Understanding Video Events : A Survey of Methods for Automatic Interpretation of Semantic Occurrences in Video," vol. 39, no. 5, pp. 489–504, 2009.