Universal Access in the Information Society manuscript No.
(will be inserted by the editor)

Behavior Monitoring for Assistive Environments using Multiple Views

Dimitrios I. Kosmopoulos

Received: date / Accepted: date

Abstract This work presents an approach to behavior understanding using multiple cameras. This approach is appropriate for monitoring people in an assistive enivronment for the purpose of issuing alerts in cases of abnormal behavior. The output of multiple classifiers is used to model and extract abnormal behaviour from both the target trajectory and the target short term activity (*i.e.*, walking, running, abrupt motion, *etc*). Spatial information is obtained after an offline camera registration using homography information. The proposed approach is verified experimentally in an indoor environment. The experiments are performed with a single moving target, however the method can be generalised to multiple moving targets, which may occlude each other, due to the use of multiple cameras.

Keywords behavior monitoring · homography · SVM · Hidden Markov Model

Mathematics Subject Classification (2000) CR I.2.10

1 Introduction

One of the key questions in creating pervasive systems for the care of the elderly is the "graceful integration with the human user" [1]. Computer vision lends itself as a very appealing method, due to the fact that it is non-intrusive. The main challenge in this case is to transform the video stream into a useful source of information. The main problems in that case are how to track people in the captured video stream, how to identify and label individuals and how to analyze their behaviors. This paper deals mainly with the third problem.

Motion analysis in video, and particularly human behaviour understanding, has attracted many researchers [2], mainly because of its fundamental applications in video surveillance, video indexing, virtual reality and computer-human interfaces. The automatic modeling and recognition of human behaviour to reduce human intervention in assistive or other environments is one of the most challenging problems in computer vision. The related systems are envisaged to automatically detect, categorize and recognize human behaviours, calling for human attention only when necessary. This is expected to increase the effectiveness of 24/7 monitoring services for elderly or patients and make such services financially viable.

The research in the area of behaviour understanding concentrates mainly on the development of methods for the analysis of visual data in order to extract and process information about the behavior of actors in a scene. Many methods have been proposed, as dicussed in the next section. The challenges are the occlusions in crowded scenes and the lack of well defined spatial information in case of 2D tracking. In systems that perform 3D tracking through multiple cameras, the high complexity poses additional performance constraints.

The goal of the method proposed in this paper is to solve the problem by modelling normal and abnormal behaviours. Multiple criteria can be considered to decide whether an observed behaviour is normal or not, such as the frequency of a specific activity, the sequence of activities and the motion patterns in a specific space.

D. I. Kosmopoulos
Institute of Informatics and Telecommunications
National Centre for Scientific Research "Demokritos"
Tel.: +30-210-6503140
Fax: +30-210-6532175
E-mail: dkosmo@iit.demokritos.gr

Therefore, a system has been implemented that considers those criteria.

To overcome problems like occlusion, view variance and high complexity, a system has been developed with multiple cameras, which capitalises on homography estimation (thus avoiding the laborious camera calibration and 3D reconstruction procedures) and views from different cameras. In this method, it is assumed that humans move on the same plane, the ground plane, and that all cameras have a common field of view. The information that is used is the object's projection on the ground, which lead to determine their position. Having extracted the sequence of positions, the trajectory in 2D is defined and used for behavior characterisation. Additionally, the problem of trajectory classification is separated from the classification of the "short term actions" that are local in a spatio-temporal sense (e.g., walking, running, abrupt motion, standing still, body motion without changing ground position). Features extracted from the blob (optical flow and speed) are used for this purpose. The features that are extracted are relatively simple, and thus computationally efficient.

Abnormalities in the established patterns for short term behavior and motion (trajectory) for a person may be indicative of problems for the monitored person. For example, people with mild cognitive impairment, often characterized by greater memory loss than normal, show increased day-to-day variability in their activity at home [3].

The presented work contributes to current research in several ways:

- The presented approach reflects two different criteria of labelling an observed behaviour as normal or abnormal, since the final decision depends on the output of two different classifiers with independent inputs: short-term behaviour information and trajectory information.
- Spatial and image information is combined to extract short term behavior, based on homography information (see section 5) which provides higher accuracy compared to pure image-based techniques[1].
- The continuous Hidden Markov Model (cHMM) framework is used as an one-class classifier to classify motion patterns in an assistive environment (see section 6).

The rest of the paper is organized as follows: section 2 presents related work and illustrates the innovative aspects of the approach proposed in this paper; section 3 provides an overview of the proposed architecture; section 4 explains briefly how the target position is computed; section 5 describes the short term behaviour representation and classification, while section 6 describes the classification for trajectories; in section 7 the experimental results are provided and finally section 8 concludes this paper.

[1] An early version of this work has been presented in [4].

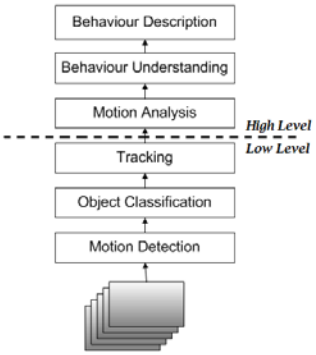

Fig. 1 The main framework for monitoring systems.

2 Related work

The various tasks for general-purpose behavior recognition in the related literature are represented by Figure 1. *Low level* tasks include such methods as motion detection, object classification and tracking. In motion detection, research is focused on either static or adaptive background subtraction or temporal differencing algorithms, aiming to isolate the foreground pixels that participate into any kind of motion observed in a given scene. Object classification is the process of classifying detected objects into such classes as humans or vehicles, appearing in a given scene.

High level processes use motion information from the low level to identify the type or nature of a moving object's activity. Motion-based techniques are mostly used for short-term activity classification (e.g., walking, running, fighting), and do not take into account object trajectories. These techniques actually calculate features of the motion itself and perform recognition of behaviours based on these features' values. Such methods have been presented by Bobick et al. in [5], where Motion Energy Images (MEIs) and Motion History Images (MHIs) are used to classify aerobic type exercises. Taking this work a step further, Weinland et al. in [6] focus on the extraction of motion descriptors analogous to MHIs, called *motion history volumes*, from multiple cameras. Then, these history volumes are classified into primitive actions. Efros et al., in [7], compute the optical flow of a given object to recognize short-term behaviours through a nearest-neighbor classification. A similar technique is followed in [8], where the motion features decide the type of action.

The recognition of motion activity patterns (trajectories) in assistive environments are of obvious importance. This has been highlighted in literature, e.g., in [9]. In the related literature there are several methods that use the target trajectory for behaviour classification using the centroid of the target blob. These methods, however, ignore the importance of what the person is doing locally (short term activity). Many of them extract trajectories in 2D images, thus having problems with view dependency and occlusions.

Hidden Markov Models are highly applicable to behavior recognition using trajectories, e.g., [10], [11], [12], [13], due to their transition and emission probabilities, to their automatically training, their simplicity, and their computational efficiency. Using HMMs motion can be viewed as piece-wise stationary signal or a short-term stationary signal. There exist several image-based techniques which model the motion at the pixel level, no matter if it results from local or global motion. In [14] the foreground pixels are clustered using fuzzy k-means clustering to model behavior patterns. The trajectories are clustered hierarchically using spatial and temporal information, and then each motion pattern is represented with a chain of Gaussian distributions. Coupled hidden Markov models were used for modeling interactions between actors, [15]. Xiang and Gong [16], use Dynamically Multi-Linked Hidden Markov Models to model actions and interactions between persons. Abstract Hidden Markov Models are used by Nguyen et al. in [17] to deal with noise and duration independence, while Wang et al. in [18] use Conditional Random Fields for behaviour recognition in order to be able to model context dependence in behaviours. In [19] a feature vector composed of features giving position and target state is used, and the behaviour representation is extracted through clustering.

The aforementioned methods seem attractive in cases of several people in the scene, however, the correspondence to real world activities is not intuitive. The approach presented in this paper differs in the sense that it decouples the position (trajectory) from the target state claiming that these can be, in many cases, separate problems, and addressing them separately may help to reduce the problem dimensionality. Moreover, in contrast to other methods, it is possibel to distinghish if the abnormality is due to abnormal trajectory or abnormal short term activity.

In behaviour understanding, few works depend on homography estimation. Park et al. in [20] use homography to extract object features, and extract behaviour by modeling people and vehicle trajectories. Ribeiro et al. in [21], estimate homography that allows the system to have an orthographic view of the ground plane to eliminate perspective distortion for a single camera. Then they calculate features in order to classify the data in four activities (Active, Inactive, Walking, Running), however no trajectory information is employed.

In the literature referenced above, in order to extract features that can be used for classification, it has been assumed that the targets move almost vertically to the camera z-axis or within a range that is small compared to the distance from the camera, so their size variation is small. Furthermore, the assumption that humans are planar objects, so that homography-based image rectification can be possible, may be true when the cameras are close to being vertical to the ground plane, e.g., cameras viewing from high ceilings, but is definitely wrong in the general case.

In this work, simple features are computed that do not require making assumptions about camera position or relative pose of target to the camera. A richer representation of an agent's behavior is also provided, by modeling both short term activity and trajectory on a 2D projection map. This representation bypasses the computationally intensive 3D world representation. Furthermore, it is closer to the human perception compared to pure image-based techniques, due to separate handling of the two information sources.

3 System Overview

This section provides an overview, thus explaining how the separate system modules cooperate to give the decision about behavior characterization. The proposed system processes video streams from several cameras with overlapping fields of view (ground plain), as displayed in figure 2.

From each camera an image sequence (video) is read, and then some low level features from the foreground objects are extracted using optical flow calculation. The foreground objects result from a background subtraction process. In parallel to this process, the pixels of each camera-specific blob are projected on the ground plane using the homography matrix, which has been calculated offline (details are provided in section 4). The maxima that result from that projection give the target position on the ground plane, from which the trajectory and the speed can be easily calculated.

The target speed, along with the view-specific features, are input to a classifier to extract short term action (as discussed in section 5). The decisions from all cameras are fused to decide the short term action as perceived by all cameras. In parallel to the aforementioned classification, the target trajectory is also classified using the currently calculated target position (details are provided in section 6). It is assumed that the

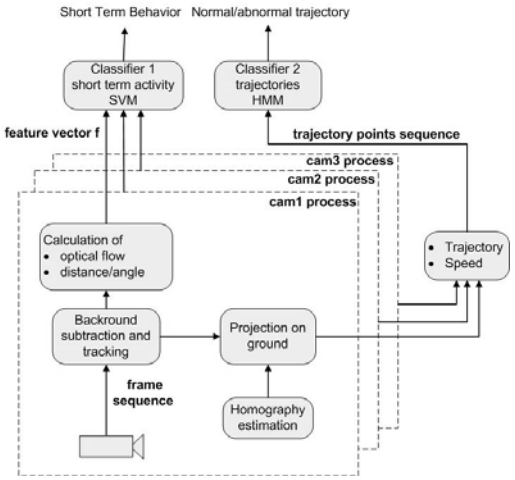

Fig. 2 The system architecture. The cam1-3 related processes and the trajectory and speed calculation correspond to the low level processing as depicted in figure 1, while the two classifiers correspond to the high level processing depicted by the same figure.

classifiers for short term action and trajectories have been trained in a supervised learning fashion. If either the short term action or the trajectory are found to be "abnormal", this is highlighted to the user.

4 Target localisation

The proposed methodology uses firstly a background subtraction method to detect motion. For the background subtraction, the adaptive Gaussian mixture background model for dynamic background modeling [22] was adopted. Similar or better methods could have been used for the same purpose without changing the overall approach, and the reader is referred to the related literature for further information.

For establishing correspondence between the world objects and the images from multiple cameras, a homography based approach is employed. Planar homographies are geometric entities whose role is to provide associations between points on different planes, the ground and the camera plane in the specific case. In the indoor environment the target moves on a ground plane, so mapping between planes is possible. In the following, it is briefly explained how the approach works.

The scene viewed by a camera comprises a predominant plane, the ground. It is assumed that a homogeneous coordinate system is attached to the ground plane, so that a point on the plane is expressed as: $\mathbf{P}_\pi=(x_{\pi 1}, x_{\pi 2}, x_{\pi 3})^T$. If this point is visible to the camera, which is a matter of proper camera configuration, the homogeneous coordinates of this point on the camera plane are given by $\mathbf{P}_c=(x_{c1}, x_{c2}, x_{c3})^T$. The homography $\mathbf{H}$ is a 3x3 matrix, which relates $\mathbf{P}_\pi$ and $\mathbf{P}_\mathrm{c}$ as follows:

$$\mathbf{P}_\pi = \mathbf{H}\cdot\mathbf{P}_\mathrm{c} \Leftrightarrow \begin{bmatrix} x_{\pi 1} \\ x_{\pi 2} \\ x_{\pi 3} \end{bmatrix} = \begin{bmatrix} h_{11}\ h_{12}\ h_{13} \\ h_{21}\ h_{22}\ h_{23} \\ h_{31}\ h_{32}\ h_{33} \end{bmatrix} \cdot \begin{bmatrix} x_{c1} \\ x_{c2} \\ x_{c3} \end{bmatrix} \quad (1)$$

Let the inhomogeneous coordinates of a pair of matching points $\mathbf{x}_c = (x_c, y_c)$ and $\mathbf{x}_\pi = (x_\pi, y_\pi)$ on the camera plane (pixel coordinates) and the ground plane correspondingly. Then:

$$x_\pi = \frac{x_{\pi 1}}{x_{\pi 3}} = \frac{h_{11}\cdot x_c + h_{12}\cdot y_c + h_{13}}{h_{31}\cdot x_c + h_{32}\cdot y_c + h_{33}} \quad (2)$$

$$y_\pi = \frac{x_{\pi 2}}{x_{\pi 3}} = \frac{h_{21}\cdot x_c + h_{22}\cdot y_c + h_{23}}{h_{31}\cdot x_c + h_{32}\cdot y_c + h_{33}} \quad (3)$$

Each point correspondence gives an equation, and four such points are sufficient for calculation of $\mathbf{H}$ up to a multiplicative factor, if no three of them are collinear. The calculation of $\mathbf{H}$ is a procedure done once offline, and normally many more points are selected to compensate for errors.

The position of each target is obtained similarly to [23]. A background subtraction algorithm extracts the silhouettes of the targets, which move on the ground plane. By replacing $(x_c, y_c, 1)^T$ and $(x_\pi, y_\pi, 1)^T$ in (1), each foreground pixel is projected on the ground plane coordinates. The projection from each camera casts a "shadow" on the ground plane, as depicted in figure 3. The maxima of those projection images indicate the positions of the monitored targets, *i.e.*, the positions where the feet touch the ground.

5 Short term activity classification

The goal of this work is to separate the classification of short term activity from trajectory classification. To this end, features are defined and extracted separately, as described in the following.

The short term behaviour refers to the type of behaviour that can be localized in a spatio-temporal sense, *i.e.*, it is brief and within a restricted space. Examples of such behaviours are walking, standing still, running, moving abruptly, *etc.*

In the present work, apart from purely image-based features as presented in the related literature, features coming from trajectory points on the common coordinate system (on the ground plane) and the optical flow are also used. The objective is to discriminate behaviors based on the target speed, e.g., walking vs. running. The features defined at the image level are able to differentiate motion types when the target does not move significantly in the global coordinate system (e.g., staying still vs. moving abruptly).

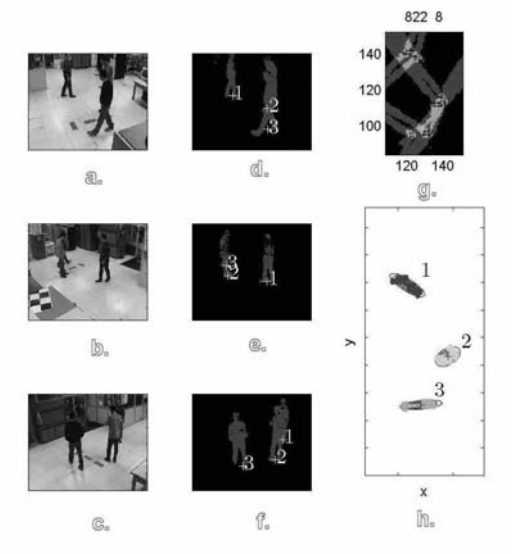

Fig. 3 View from three cameras and extraction of projections on ground plane. In a,b,c the views, in d, e, f the points extracted from background subtraction (several occlusions are present), in g the maxima of the accumulator corresponding to feet positions (represented as crosses) and in h the positions in ground plane.

The short term activity is represented here by a three - dimensional feature vector, as follows:

$$\mathrm{f(t)} = (\mathrm{v_t}, \mathrm{mean(F_t)}, \mathrm{stdev(F_t)}) \tag{4}$$

where v_t is the velocity norm on the ground plane, and F_t is the norm of the optical flow for two consecutive frames. The optical flow can be calculated using standard methods such as [24]. Based on the F_t the mean and the standard deviation are extracted in the foreground region that composes the target, which is defined after a background substraction and morphological operations to eliminate noise and holes in the related silhouette.

The multicamera approach here consists in the fusion of those vectors from multiple caemeras. A very simple yet effective fusion scheme is used, which consists in averaging the vectors defined in equation (4) over all available cameras.

The speed calculation is computationally inexpensive. The mean optical flow requires more computational time, but it can be calculated accurately in real time by limiting its calculation in the foreground regions.

Subsequently, a standard classifier can be trained to automatically classify the short term behavior to normal or abnormal using the above extracted motion information within a small time window, thus ensuring real time performance. Such a classifier is a Support Vector Machine (SVM) [25].

6 Trajectory classification

In order to classify trajectory, classifiers able to handle time series may be used. One of the most popular ones is the continuous Hidden Markov Model. In the specific case, each (x,y) ground position is supposed to be the observation vector. Given a continuous Hidden Markov Model, which models each observation as a mixture of distributions, e.g., Gaussians, it is possible to classify in real time the trajectory up to the current moment as normal or abnormal, based on a set of training parameters which have been learnt offline by using an EM algorithm see, e.g.,[26]. The problem is stated as follows: if the forward variable

$$a_t(i) = P(O_1 ... O_t, q_t = S_i|\lambda) \tag{5}$$

is the probability of the partial observation sequence O_1 $...O_t$ and state S_i at time t given the model λ, then the $a_t(i)$ is calculated inductively by the following:

$$\alpha_1(i) = \pi_i b_i(O_1), 1 \leq i \leq N \tag{6}$$

where N is the number of states and π_i the state priors, $b_i(O_1)$ the probability of observation O_1 at t=1, given that the current state is i. Furthermore,

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_j(O_{t+1}), 1 \leq t \leq T-1, 1 \leq j \leq N \tag{7}$$

where a_{ij} is the transition probability from state i to state j, and $b_j(O_{t+1})$ is the probability of observation O_{t+1} at $t+1$ given that the current state is j. Then the desired probability is the sum of terminal probabilities:

$$P(O_1, ...O_T|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{8}$$

The observation probability is given by:

$$b_j(\mathbf{O}) = \sum_{m=1}^{M} c_{jm} N(\mathbf{O}, \mu_{jm}, \Sigma_{jm}) \tag{9}$$

where c_{jm} is the probability that the sample is drawn from the m-component and μ_{jm} and Σ_{jm} are the mean vector and the covariance matrix of the m-component in the j-state.

When the value calculated by equation (8) is very small, then the sequence can be classified as abnormal. The problem faced by such an approach is the variable length of the sequences. It has been shown in [26] that the likelihood becomes smaller as the sequence gets longer, and thus comparable to the likelihood of abnormal trajectories. Furthermore, even double precision is not enough to support the reestimation procedure of the $a_t(i)$, π, $b_j(O_t)$ for long sequences (e.g., length 100). A solution is provided in [26], where a scaling scheme is proposed. The same scheme has been applied in the presented method.

7 Experiments

The experiments were performed in an in-house laboratory, where three cameras with overlapping fields of view have been installed. The hypothetical scenario concerned patients that were supposed to follow certain paths (like from bedroom to bathroom, kitchen to hall, *etc.*) within certain tolerances. This can represent a sequence of approximate positions which are normally accessed in a certain order, but in a non deterministic fashion. Deviation from those normal patterns might be indicative of health problems or a panic situation, e.g., when moving from one room to another one in a zig-zag manner instead of following a straight trajectory.

The employed cameras were the AXIS 214 PTZ (network cameras). The frames from the camera were received through http requests. The communication with the cameras was performed through an IP network. For frame synchronization, an NTP server was used, which gave time stamps to each frame, so the closest frame triplet was examined in each processing cycle. The phenomenon of losing frames due to processing or network delays (a non dedicated LAN was used) was common, and led to a frame rate of 10-20 frames per second.

To evaluate the system, 34 video shots have been used, with several behaviors of approximately 170000 frames per camera. Four actors have been involved for this purpose performing specific actions during each video.

7.1 Short term behavior

For the task of short term behavior classification a binary SVM with polynomial kernel was used. More specifically, the SVM light library [27] was employed for this purpose.

The training for the short term behavior recognition has been very time consuming, because the frames representing certain behaviors have to be manually selected after visual examination of the captured frames. It is possible that an abnormal video shot produces several feature vectors that correspond to normal behaviors, which if not excluded can corrupt the training process. The *normal* short term activity corresponds to observed actions like "walking", "staying still", "active". The *abnormal* short term sctivity corresponds to observed actions like "running" or "abrupt motion", which are very rare to be observed in an assistive environment. Several hundred vectors from each of the above behaviors have been employed for training each class.

Due to the fact that there are several noisy observations, the obtained results were filtered not by considering a single classification result, but by voting within a time window (approximately half a second). This duration can cope with noise and at the same time preserve the abnormalities, for example due to abrupt motion, which are generally of short duration.

To clarify the benefit of using multiple cameras as opposed to single view approaches that have been already presented in the related literature, the obtained results from each view have been compared with the ones from all cameras. For the classification using single views, the speed information has been replaced with blob centroid speed in the image (pure image-based technique). The results can be seen in Table 1. In the same table, given that TP are the true positives, TN the true negatives, FP the false positives, FN the false negatives:

$RN = TP/(TP + FN)$
$PN = TP/(TP + FP)$
$RA = TN/(TN + FP)$
$PA = TN/(TN + FN)$

As expected, the benefit for employing spatial information is clear. What is not visible in the table is that the field of view is expanded and that with multiple cameras one can still characterize behaviors even if the target is lost from one camera.

The processing was perfomed at frame rates (real time), which was expected because of the features used, which were quite simple. Approximately, 70% of the processing power was consumed for background subtraction.

The classification errors are mainly due to following factors:

- Problems in the background subtraction. The employed method just models single pixels without considering the neighboring pixels (texture) or the agent motion. As a result, agents wearing colors similar to the background are considered as background. By employing more elaborate background subtraction methods, this error source could be minimised but this would also increase the processing requirements.
- The optical flow which is used as the main feature does not perform equally well for silhouettes of high and low resolution. The consequence is that close views would give significantly higher optical flow, even for normal motion, than views taken far from the agent. This is now partially compensated by taking multiple views, but is expected to be even better handled by considering both distance from the camera and angle of motion relative to the camera.

Table 1 Short term behavior classification results using single camera (for all three cameras) and multiple views

	camera1	camera2	camera3	all cameras
PN	0.765	0.902	0.785	0.922
RN	0.824	0.768	0.674	0.874
PA	0.821	0.831	0.734	0.867
RN	0.802	0.841	0.755	0.852

Table 2 Trajectory classification results using single camera (for all three cameras) and multiple views

	camera1	camera2	all cameras
PN	0.681	0.754	0.783
RN	0.722	0.798	0.821
PA	0.832	0.889	0.931
RN	0.852	0.871	0.912

7.2 Trajectory classification

For the task of trajectory classification, the Hidden Markov Model Toolbox for Matlab [28] was adopted. The HMM was trained using only normal trajectories, which represent the ordinary motion patterns observed in the monitored space. A total database of 98 trajectories was used.

Similarly to the previous subsection, the results obtained from each view were compared with the ones from all cameras. For single cameras the trajectory of the target blob centroid in the image coordinates is used, as opposed to the spatial coordinates when considering all views.

A five-fold cross validation method has been applied, *i.e.*, five sequences have been evaluated, while the remaining normal ones have been trained by applying all possible combinations. The calculation of the log likelihood has been performed using the scaling factor described in section 6 to compensate for low log likelihoods in normal long sequences. The abnormal behaviors were characterized by low log likelihood. The results using a continuous HMM with five states and two mixture components for trajectory analysis are reported in Table 2.

The allowed tolerances in the trajectories at runtime are directly associated to the learned observation model, which is represented by a Gaussian mixture model. The high variability of trajectories during training results in higher standard deviation of the mixture components, and thus in higher tolerance to trajectory ambiguity.

It is worth mentioning that parts of the trajectories were not visible in most sequences in camera 3, so no results are provided for that camera. Not being able to monitor the whole trajectory, and thus to provide meaningful results, is not uncommon in monitoring systems and underlines the benefits of using multiple cameras to cover wider areas.

The trajectory classification was performed in real time. Whenever the agent was deviating from a trajectory characterised as normal, the system was able to highlight it immediately.

Errors in the trajectory classification may result from wrong position exctraction, caused by background extraction errors. Provided that the target silhouette is correctly extracted, the error in position was less than 0.2m, while in the opposite case it could be 0.5-0.7m. A low pass filtering in the position sequence helped to alleviate many of those errors.

8 Conclusions

This paper has presented a methodology for modeling and understanding human behavior using multiple views in real time. The proposed methodology decouples the short term activity and the trajectory classification problems, thus reducing the problem dimensionality. Spatial information is exploited to extract short term behavior, bypassing the computationally intensive 3D world representation. Furthermore, the presented model is closer to human perception compared to pure image-based techniques, due to separate handling of the two information sources.

The experiments were performed with a single moving target, though the method can be generalised to multiple moving targets, which may occlude each other. The multiple cameras allow for efficient handling of occlusions. One of the underlying goals is to use techniques such as the one described in [29] to extend the framework for multiple targets. Despite the fact that the features used are quite simple, and the fusion scheme is simple as well, overall approach has been shown to be feasible, and to provide promising results. The use of multiple cameras extends the limited field of view and provides higher accuracy compared to single-view approaches. The performance of the presented method is dependent on the success of the background subtraction method. Thus, the more the background subtraction methods evolve, the better the system performance will be.

The planned next steps include the the use of more complex features for better representation of relative position to the camera, as well as more complex fusion schemes with the ultimate goal of behavior recognition for multiple persons under occlusions. As proposed in [3], in the future it will be possible to use behavioral

monitoring to provide more precise information, such as daily reporting of unexpected variations in behaviors: Did the patient not sleep well? Did the patient fail to eat? These reports can help caregivers focus on providing the most effective care rather than on monitoring patient activities.

References

1. Stanford, V.: Using pervasive computing to deliver elder care. Pervasive Computing, IEEE **1**(1), 10–13 (Jan-Mar 2002). DOI 10.1109/MPRV.2002.993139
2. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. **104**(2), 90–126 (2006). DOI http://dx.doi.org/10.1016/j.cviu.2006.08.002
3. Hayes, T.L., Pavel, M., Larimer, N., Tsay, I.A., Nutt, J., Adami, A.G.: Distributed healthcare: Simultaneous assessment of multiple individuals. Pervasive Computing, IEEE **6**(1), 36–43 (Jan.-March 2007). DOI 10.1109/MPRV.2007.9
4. Kosmopoulos, D., Antonakaki, P., Valasoulis, K., Katsoulas, D.: Monitoring human behavior in an assistive environment using multiple views. In: PETRA '08: Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments, pp. 1–6. ACM, New York, NY, USA (2008). DOI http://doi.acm.org/10.1145/1389586.1389624
5. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**(3), 257–267 (2001). DOI http://dx.doi.org/10.1109/34.910878
6. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Comput. Vis. Image Underst. **104**(2), 249–257 (2006). DOI http://dx.doi.org/10.1016/j.cviu.2006.07.013
7. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: IEEE International Conference on Computer Vision, pp. 726–733. Nice, France (2003)
8. Hauptmann, A., Gao, J., Yan, R., Qi, Y., Yang, J., Wactlar, H.: Automated analysis of nursing home observations. Pervasive Computing, IEEE **3**(2), 15–21 (April-June 2004). DOI 10.1109/MPRV.2004.1316813
9. Chan, M., Campo, E., Esteve, D.: Monitoring elderly people using a multisensor system. In: D. Zhang, M. Mokhtari (eds.) Toward a Human-friendly Assistive Environment: ICOST '2004, 2nd International Conference on Smart Homes and Health Telematics, pp. 162–169. IOS Press (2004)
10. Bregler, C., Malik, J.: Learning appearance based models: Mixtures of second moment experts. In: M.C. Mozer, M.I. Jordan, T. Petsche (eds.) Advances in Neural Information Processing Systems, vol. 9, p. 845. The MIT Press (1997). URL citeseer.ist.psu.edu/article/bregler97learning.html
11. Ivanov, Y.A., Bobick, A.F.: Recognition of visual activities and interactions by stochastic parsing. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 852–872 (2000). DOI http://dx.doi.org/10.1109/34.868686
12. Bashir, F.I., Qu, W., Khokhar, A.A., Schonfeld, D.: Hmm-based motion recognition system using segmented pca. In: ICIP (3), pp. 1288–1291 (2005)
13. Sukthankar, G., Sycara, K.: Robust recognition of physical team behaviors using spatio-temporal models. In: In Proceedings of Workshop on Modeling Others from Observations (MOO 2005). (2006)
14. Hu, W., Xiao, X., Fu, Z., Xie, D., Tan, T., Maybank, S.: A system for learning statistical motion patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(9), 1450–1464 (2006). DOI http://doi.ieeecomputersociety.org/10.1109/TPAMI.2006.176
15. Brand, M., Oliver, N., Pentland, A.: Coupled hidden markov models for complex action recognition. In: CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), p. 994. IEEE Computer Society, Washington, DC, USA (1997)
16. Xiang, T., Gong, S.: Beyond tracking: Modelling activity and understanding behaviour. Int. J. Comput. Vision **67**(1), 21–51 (2006)
17. Nguyen, N., Bui, H., Venkatesh, S., West, G.: Recognising and monitoring highlevel behaviours in complex spatial environments. In: In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR-03) (2003)
18. Wang, T., Li, J., Diao, Q., Hu, W., Zhang, Y., Dulong, C.: Semantic event detection using conditional random fields. In: CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop, p. 109. IEEE Computer Society, Washington, DC, USA (2006). DOI http://dx.doi.org/10.1109/CVPRW.2006.190
19. Xiang, T., Gong, S.: Unsupervised video behaviour profiling for on-the-fly anomaly detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2008)
20. Park, S., Trivedi, M.M.: Analysis and query of person-vehicle interactions in homography domain. In: VSSN '06: Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks, pp. 101–110. ACM, New York, NY, USA (2006). DOI http://doi.acm.org/10.1145/1178782.1178798
21. Ribeiro, P., Santos-Victor, J.: Human activities recognition from video: modelling, feature selection and classification architecture. In: Proc. Workshop on Human Activity Recognition and Modelling (HAREM 2005), pp. 61–70 (2005)
22. Zivkovic, Z., van der Heijden, F.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recognition Letters **27**(7), 773–780 (2006)
23. Khan, S.M., Shah, M.: A multiview approach to tracking people in crowded scenes using a planar homography constraint. In: ECCV (4), pp. 133–146 (2006)
24. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: 7th International Joint Conference on Artificial Intelligence (IJCAI), pp. 674–679 (1981)
25. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer (1995). URL http://svmlight.joachims.org
26. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989). URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=18626
27. Joachims, T.: Svm light, support vector machine (2008). URL http://svmlight.joachims.org
28. Murphy, K.: Hidden markov model (hmm) toolbox for matlab (2005). URL http://www.cs.ubc.ca/ murphyk/Software/HMM/hmm.html
29. Kesidis, A.L., Kosmopoulos, D.I.: Robust occlusion handling with multiple cameras using a homography constraint. In: International conference on Computer Vision theory and Applications (2), pp. 560–565 (2009)