ABSTRACT

This paper presents our approach in understanding the behavior of humans moving on a plane using multiple cameras. This approach is appropriate for monitoring people in an assistive environment for the purpose of issuing alerts in cases of abnormal behavior. We perform camera registration based on homography estimation and we extract position on 2D projection map. We use the output of multiple classifiers to model and extract abnormal behavior from both the target trajectory and the target short term activity (i.e., walking, running, abrupt motion etc). 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.

Categories and Subject Descriptors
I.4.9 [Image Processing and Computer Vision]: Applications—video analysis; I.2.10 [Artificial Intelligence]: Vision and Scene Understanding—motion; 1.5.4 [Pattern Recognition]: Applications—computer vision

Keywords
behavior monitoring, homography, optical flow, Hidden Markov Model

1. INTRODUCTION

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions Dept, ACM, Inc., fax +1 (212) 869-0481, or e-mail permissions@acm.org.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
2. RELATED WORK

Abnormalities in the established patterns for short term behavior and 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 [7].

The rest of the paper is organized as follows: in section 2 we present the previous work done, which will help to evaluate the innovations of the proposed approach; section 3 provides an overview of the proposed architecture; section 4 explains briefly how we compute the principal axis and how this is projected on the ground plane; section 5 describes the short term behaviour representation and classification, while section 6 describes the classification for trajectories; in section 7 we provide the experimental results and section 8 concludes this paper.

2. RELATED WORK

The framework for monitoring systems in the related literature is divided in layers as displayed in Fig. 1. The first low level tasks such as motion segmentation, object classification and tracking are not going to be developed in this work. The tracker provides the system with verified data referring human motion, and we present methods that take a step further and analyze that motion providing behaviour description. In our work, we not only focus on the higher level of behaviour description, but also on fusion of data from many cameras, trying to overcome some of the difficulties discussed before, such as occlusion and view dependence. As mentioned earlier, we examine both the type of motion (or short term activity) and the target trajectory to model behavior and detect abnormal events. In the related literature these two are mostly treated separately.

Motion based techniques are normally used for short term action classification, without considering the target trajectories. These calculate features of the motion itself and perform recognition of behaviors based on them. There are many methods for this kind of recognition, for example Bobick et al. in [2] use Motion Energy Images (MEIs) and Motion History Images (MHIs) to classify "aerobic" - type exercises. Taking this work another step further Weinland et al. in [20], focus on the extraction of motion descriptors analogous to MHIs, called motion history volumes, from multiple cameras and their classification into primitive actions. In [5], compute optical flow of the object of interest to recognize short term behaviours (e.g., walking, running, fighting) in a nearest neighbour framework. Similar technique is followed in [6] where the motion features decide the type of action.

There are several other 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 (short term activity). They also extract trajectories in 2D images thus having problems with view dependency, occlusions etc.

Hidden Markov Models are highly applicable to behavior recognition using trajectories, e.g., [4], [10], [1], [18], due to their transition and emission probabilities, to their automatically training, their simplicity, their computational efficiency and mainly because motion can be viewed as piecewise stationary signal or a short-term stationary signal. There exist several image-based techniques, which model the motion in the pixel level no matter if it results from local or global motion. In [9] 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, [3]. Xiang and Gong [21], use Densely Multi-Linked Hidden Markov Models to model actions and interactions between persons. Abstract Hidden Markov Models are used by Nguyen et al. in [13] to deal with noise and duration independence, while Wang et al. in [19] use Conditional Random Fields for behaviour recognition in order to be able to model context dependence in behaviours. In [22] a feature vector composed of features giving position and target state is used and the behaviours’ representations are extracted through clustering.

Figure 1: The main framework for monitoring systems.
The aforementioned methods seem attractive in cases of several people in the scene, however, the correspondence to real world activities is not intuitive. Our approach differs in the sense that we decouple 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 those methods we are able to say if the abnormality is due to abnormal trajectory or abnormal short term activity.

In behaviour understanding, few works are depending in homography. Park et al. in [14] use homography to extract object features, and with spatio-temporal relationships between people and vehicle tracks to extract semantic information from interactions calculated from relative positions. Ribeiro et al. in [16], 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 common plane, e.g., cameras viewing from high ceilings, but is definitely wrong in the general case.

In our work we compute simple features that work without making assumptions about camera position or relative pose of target to the camera. We also provide a richer representation of an agent’s behavior 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.

\section{3. SYSTEM OVERVIEW}

The proposed system processes video streams from several cameras with overlapping fields of view (ground plain) as displayed in Fig.2. From each camera we take an image sequence (video) and then we extract some low level features from the foreground objects (resulting from optical flow). The foreground objects result from a background subtraction process. In parallel to this process we project on the ground plane the principal axis of each camera-specific blob using the homography matrix, which has been calculated offline. 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 perceived by each camera separately. The decisions from all cameras are fused to decide the short term action as perceived by all cameras. In parallel to the aforementioned classification we also classify the target trajectory using the currently calculated target position. It is assumed that the 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 by the “anomaly detection” module.

\section{4. TARGET LOCALISATION VIA HOMOGRAPHIES}

The (planar) homographies are geometric entities whose role is to provide associations between points on different planes. Assume that the scene viewed by a camera comprises a predominant plane, the ground for example. Assume as well that a coordinate system is attached, so that a point on the plane is expressed as: $P_\pi = (X,Y,1)^T$. Finally, assume that the coordinates of this point on the camera plane is $P_c = (x,y,1)^T$. The homography $H$ is a 3x3 matrix which relates $P_\pi$ and $P_c$ as follows:

$$P_\pi = H \cdot P_c \quad (1)$$

We assume that the target moves on a predominant ground plane. Multiple cameras view the target. The homographies between the cameras and the views project subject information to the ground plane. Synergy of information from all views into the ground plane allows for increased accuracy in the localization of the subject. Our approach is inspired by [11]. In that work, the homographies between each view and the ground plane are calculated. Subsequently, foreground likelihoods in all views are computed. The pixels belonging to the foreground blobs are projected to the ground plane using (1). The projected likelihood maps are multiplied, and a synergy map is applied. The maxima of the map correspond to ground point position of the viewed target on the ground plane, that is, the position where its feet touch the ground. The weak point of that method is that it requires that the foreground blobs comprise the points where the subject touches the ground. Unfortunately, this is not always the case due to image noise. Moreover, due to noise the extracted foreground blobs do not accurately correspond to the viewed subject. Using the principal axis of a subject instead of a foreground likelihood map increases the robustness against noise: Since foreground pixels corresponding to a person are in general symmetrically distributed along the principal axis, the errors of monocular motion subtraction are also symmetrically distributed along the axis [8]. Furthermore, the principal axis of the target will always contain the ground plane of the target: it will be the axis intersection with the ground plane. Hence, we employ principal axes instead of the foreground map as feature for subject localization. More specifically, the principal axis of the target in each view is extracted as in [8]. A two-dimensional accumulator is attached on the ground plane whose role is similar to the synergy map of [11]. The axes of the views are projected to the accumulator using (1). The maximum of the accumulator corresponds to the position of the subject on the ground plane. Figure 3 illustrates the output of our approach. In a-c we show the localized principal axes in each view. In d the contents of the accumulator after the projection of the axes on the ground plane. The maximum is the position of the subject. This position is back-projected to each view and highlighted with a '*'. The target position computed by our approach is considerably more accurate than using only one camera for position determination, since it comes as the outcome of fusion from all cameras. Furthermore, it is more accurate than integrating multiple camera information using foreground maps, since information of the ground position of the subject is always delivered from all
Figure 2: View from three cameras and extraction of principal axis projection on ground plane from two of them. In c the projection is not visible, however, the corresponding accumulator is still created in d. In d three accumulators are visible - two of them very “close” to each other.

cameras. This is the case for figure 3 in our experiments.

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 we define and extract features separately as described in the following. The short term activity is defined as the activity that takes place locally in the global coordinate system and within a short time period, e.g., a few frames. It is associated with types of motion like walking, running, standing still, body motion while no translation takes place on the global coordinate system, abrupt motion. In the present work, we use features coming from trajectory points on the common coordinate system (on the ground plane) and the optical flow. The former provide the target speed (walking vs running), while the latter are able to differentiate motion types when the target does not move significantly in the global coordinate system (staying still vs moving abruptly).

The short term activity is represented by a 3-dimensional feature vector, as follows:

\[
f(t) = (v_t, \text{mean}(F_t), \text{std}(F_t))
\]

where \(v_t\) is the velocity norm on the ground plane respectively and \(F_t\) is the norm of the optical flow for two consecutive frames. This measure is averaged 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.

In the following 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 an SVM, which was used in our experiments as will be mentioned in section 7.

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 our 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.,[15]. The problem is stated as follows: if the forward variable

\[
a_t(i) = P(O_t ... O_N, q_t = S_i/\lambda)
\]

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

\[
a_t(i) = \pi_i b_i(O_t), 1 \leq i \leq N
\]

where \(N\) is the number of states and \(\pi_i\) the state priors, \(b_i(O_t)\) the probability of observation \(O_t\) at \(t=1\), given that we are in state \(i\)

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

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

\[
P(O_1, ... O_T|\lambda) = \sum_{i=1}^{N} \alpha_T(i)
\]

The observation probability is given by:

\[
b_j(O) = \sum_{m=1}^{M} c_{jmN}(O, \mu_{jm}, \Sigma_{jm})
\]

where \(c_{jm}\) the probability that the sample is drawn from the \(m\)-component and \(mu_{jm}\) and \(\Sigma_{jm}\) the mean vector and the covariance matrix of the \(m\)-component in the \(j\)-state.

7. EXPERIMENTS

For our experiments we have used our lab, where we have installed three cameras with overlapping filed of view. We have simulated a room, where the patients are supposed to follow a certain path within certain tolerance. This can represent a sequence of positions which are normally accessed at a certain order in a non deterministic fashion. Deviation from that normal pattern of activity might be indicative of health problems. To evaluate our system we have used 34 sequences with several behaviors of approximately 170000 frames per camera. Four actors have been used for this purpose performing specific actions during each video. This can be problematic because it is impossible to perform an action in a consistent way through a specific duration and therefore we have empirically excluded a part of the automatically produced vectors in order to
help class separation. The average results using a continuous HMM for trajectory analysis are given in Table 1. We applied a five-fold validation method, i.e., 5 sequences have been evaluated while we have trained with the rest ones by applying all possible combinations. trained with 20 and 40 states, where TP are the true positives, TN the true negatives, FP the false positives, FN the false negatives. $PN = TP/(TP + FN)$, $RN = TP/(TP + FP)$, $RA = TN/(TN + FP)$, $PA = TN/(TN + FN)$. The log likelihood has been normalised according to trajectory duration and a threshold was experimentally set in order to find outliers.

A similar cross validation method has been used for the short term - local behaviors, where we have used a binary SVM with polynomial kernel. Furthermore, the "normal" short term activity has to be something like "walking", "staying still" or "active" and the "abnormal" refers to "running" or "abrupt motion". The results are given in Table 2. The results are meaningful and promising. However, we have performed some initial experiments in which the relative pose of the target to each camera was considered (calculated using the homography information) and thus we could possibly provide better results. This is due to the fact that the motion vertical to the camera axis provided higher optical flow compared to motion towards to or away from the camera. This effect is now only partially compensated by the fact that the cameras view the target from different viewpoints.

### 8. CONCLUSIONS

In our work we have provided a richer representation of an agent’s behavior compared to most current methods by modeling both short term activity and trajectory on a 2D projection map. Through the proposed representation scheme we have bypassed the computationally intensive 3D world representation. Furthermore, our model is closer to the 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, due to the use of multiple cameras. Despite the fact that the features that we used are quite simple and the fusion scheme is simple as well the concept seems to work and provides very promising results.

Our next steps include the monitoring of multiple persons and the use of more complex fusion schemes. As proposed in [7] in the future we can use behavioral monitoring to provide more acute 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.

### 9. ACKNOWLEDGEMENT

This research has been supported by the Greek Secretariat of Research and Technology, grant number EHG-91, project SemVeillance (www.semveillance.gr).

### 10. REFERENCES


### Table 1: Evaluation of trajectory classification

<table>
<thead>
<tr>
<th>States</th>
<th>PN</th>
<th>RN</th>
<th>PA</th>
<th>RA</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>78.34</td>
<td>82.14</td>
<td>93.16</td>
<td>91.24</td>
</tr>
<tr>
<td>40</td>
<td>83.53</td>
<td>89.29</td>
<td>94.74</td>
<td>91.32</td>
</tr>
</tbody>
</table>

### Table 2: Evaluation of local behavior classification

<table>
<thead>
<tr>
<th></th>
<th>PN</th>
<th>RN</th>
<th>PA</th>
<th>RA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>92.24</td>
<td>87.42</td>
<td>86.76</td>
<td>85.24</td>
</tr>
</tbody>
</table>


