Vision-Based Detection of Unusual Patient Activity

Paulo Vinicius Koerich Borges [a,1], Navid Nourani-Vatani [a,b,2]
[a] *Autonomous Systems Laboratory - CSIRO ICT Centre*
[b] *University of Queensland*

Abstract. Automated patient monitoring in hospital environments has gained increased attention in the last decade. An important problem is that of behaviour analysis of psychiatric patients, where adequate monitoring can minimize the risk of harm to hospital staff, property and to the patients themselves. For this task, we perform a preliminary investigation on visual-based patient monitoring using surveillance cameras. The proposed method uses statistics of optical flow vectors extracted from the patient movements to identify dangerous behavior. In addition, the method also performs foreground segmentation followed by blob tracking in order to extract shape and temporal characteristics of blobs. Dangerous behaviour includes attempting to break out of safe-rooms, self-harm and fighting. The features considered include a temporal and multi-resolution analysis of blob coarseness, blob area, movement speed and position in the room. This information can also be used to normalize the other features according to estimated position of the patient in the room. In this preliminary study, experiments in a real hospital scenario illustrate the potential applicability of the method.

Keywords. Patient Monitoring, Computer Vision.

Introduction

Effective monitoring of psychiatric patients has always been a major concern in hospitals. In this work we propose and investigate the applicability of a patient monitoring method targeting “safe-rooms” in emergency departments. In such scenarios, patients can often become aggressive due to drug abuse or mental diseases. They can also attempt escapes as well as cause harm to property and to themselves. A possible monitoring alternative is the employment of traditional surveillance cameras combined with audio for the automated detection of “loud” events. The use of audio, however, conflicts with current privacy legislations.

For these reasons, we propose a vision-based method for the recognition of unusual human action, such at attempting to break out, fighting with hospital staff and self-harm, for example. The method uses video data from surveillance cameras and analyses the

[1] The author is with the Autonomous System Laboratory, ICT Centre, CSIRO. Adress: 1, Technology Court, Pullenvale, QLD, 4069, Australia. paulo.borges@csiro.au, vini@ieee.org

[2] The author is with [1] and also with the School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, QLD, 4072, Australia. navid.nourani@csiro.au

statistics of vectors extracted from optical flow information. Each vector, with its position, direction and magnitude carries information that can be observed locally and globally in each frame. We argue that the magnitudes and coherency among vectors can be efficient metrics to detect some types of unusual behavior, in particular those related to fast movements.

Apart from the optical flow, we also combine traditional foreground segmentation followed by blob tracking in order to extract shape and temporal characteristics of blobs. These characteristics include blob coarseness, area, speed and position in the room. The latter feature can be determined when the homography transformation between the camera and ground plane is known. We test the method using video data extracted from a real hospital environment, and the results indicated the applicability of the proposed technique.

1. Related Work

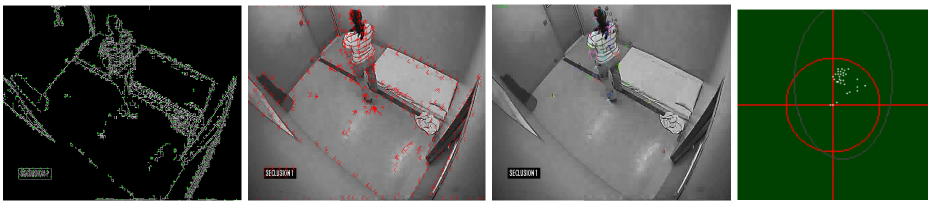
In the past few years, a wide range of approaches have been proposed in the literature on the topic of recognition of human actions by analyzing different types of visual information. Among these methods, spatio-temporal interest points (STIPs) [1] and silhouette analysis have been arguably the most commonly used visual features. In this section we provide a brief review on the work of activity detection/classification based on silhouettes and interest points. For a more comprehensive review on alternative types of analysis, such as those based on skeletons, for example, the reader is pointed to [1,2,3].

The usage of silhouettes to recognize and classify specific human actions generally considers the assumption that human movement can be seen as a continuous progression of the body posture and is essentially based on the background segmentation algorithms [4] applied to video surveillance. Although in outdoor scenarios or low-resolution images silhouettes can be very noisy, in controlled environments (such as indoors) silhouettes are often able to efficiently indicate the shape of body poses.

Silhouette-based activity classification can be broadly divided in two main classes. The first class consists of extracting action descriptors from a sequence of silhouettes in consecutive frames, such that traditional classifiers can be employed for the recognition. Using this approach, the action descriptors capture and combine both temporal and spatial behavior characteristics of the activity.

A common technique is to accumulate silhouettes to generate motion energy images (MEI) as well as motion history images (MHI) [5]. Hu moments [6] can be extracted from both MEI and MHI to serve as action descriptors, and action classification is based on the Mahalanobis distance between each moment descriptor of the known actions and the input one. Chen et al. [7] proposed a method in which star figure models are fitted to human silhouettes aiming at identifying the five extremities of the shape that correspond to the head, arms, and legs. To model the spatial distribution of the five points over time, Gaussian mixture models were used, ignoring the temporal order of the silhouettes in the action sequence.

The other main class of methods is to extract characteristics from each silhouette and to create a dynamic model of the action of interest. These approaches frequently employ statistics-based techniques such as conditional random fields (CRF) and hidden Markov models (HMM). The feature extracted from each silhouette captures the shape



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

Figure 1. Optical flow analysis. Figure (a) illustrates the interest points using Harris corner detection. Figure (b) shows raw optical flow vectors whereas Figure (c) shows the corresponding vectors after filtering very small (noise-like) vectors. Figure (d) shows a polar plot of optical flow activity, where each white dot represents a vector with the corresponding angle and magnitude.

of the body and possible local motion. Such a methodology using HMM to model the temporal behavior was proposed by Divis and Tyagi [8].

Similarly to this paper, several works perform activity analysis aiming at people well-being. To detect falls, for example, Nater et al. [9] use a hierarchical approach based on silhouettes, whereas Li et al. [10] extend the silhouette approach to 3D sensors.

In contrast to the methods above, we approach the problem by combining optical flow analysis with the silhouette based analysis, focusing the application on hospital safe-rooms, as discussed in the next section.

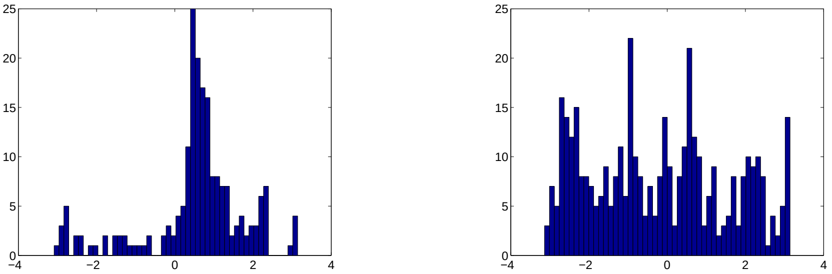
2. Proposed Method

The proposed solution is divided in two main parts, combining the analysis of optical flow and tracked blobs.

2.1. Optical Flow Analysis

For the optical flow, we extract sparse flows information from the video sequence, such that each flow is represented by a vector. The method uses Harris corner detection [11] for the selection of feature points. For each frame i, a set of K vectors is obtained. Each vector carries information about its position, direction and magnitude. Hence, for every frame a matrix of features $\mathbf{F}$ can be written as $\mathbf{F} = [\mathbf{x}\,\mathbf{y}\,\boldsymbol{\theta}\,\mathbf{r}]^\top$, where $\mathbf{x}$, $\mathbf{y}$, $\boldsymbol{\theta}$ and $\mathbf{r}$ correspond to the horizontal position, vertical position, direction, and magnitude of each flow, respectively. Each vector in $\mathbf{F}$ is given by $\mathbf{x} = [x_1 \cdots x_K]$. Similar notation is used for the other elements ($\mathbf{y}$, $\boldsymbol{\theta}$ and $\mathbf{r}$) in $\mathbf{F}$.

Observing the statistics of each feature, information about the current action can be inferred. In particular, we perform a histogram analysis of the features, and how they evolve from frame to frame. These statistics include the mean, variance, skewness and kurtosis. As a simple example, actions like leaving or entering the room present an approximately unimodal histogram with $\mu_\mathbf{x}$ and $\mu_\mathbf{y}$ on the center of the door region, where $\mu_\mathbf{x}$ and $\mu_\mathbf{y}$ represent the means of $\mathbf{x}$ and $\mathbf{y}$, respectively. In this case, the vector directions are also generally coherent. In the case of a fight, on the other hand, the flow vectors present a random behavior, both in term of direction and magnitude, as some parts of the body move much faster than others. For illustration, the histograms in Figures 2a and 2b



(a) Histogram of the angles of a ‘coherent’ movement.

(b) Histogram of the angles of a ‘non-coherent’ action.

Figure 2. Histograms illustrating the distribution of the angles (between $-\pi$ and $+\pi$) of the optical flow vectors for ‘entering the room’ (a) and ‘fighting’ (b), respectively.

show the distribution of the angles of the optical flow vectors for ‘entering the room’ and ‘fighting,’ respectively.

2.2. Analysis of Tracked Blobs

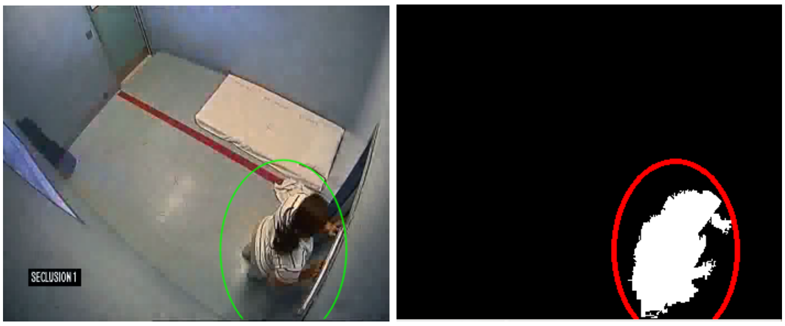
In the following we describe the tracking algorithm used, which can be divided into three main stages: background segmentation, foreground blob detection, and blob tracking.

2.2.1. Background Segmentation

In its simplest form, background segmentation can be achieved by averaging a number of frames to generate the model and subsequently labeling as foreground any pixel that exceeds a given difference threshold. For more robustness and ability to deal with shadows, however, a more sophisticated algorithm is necessary for the segmentation. In our case, we use the algorithm proposed by Li et al. [12], which has the capability of processing relatively complex backgrounds. The technique is based on pixel colour and co-occurrence statistics. The pixel colour and the color co-occurence distributions are represented by histograms, which serve as basis for the Bayes decision rule to classify pixels to foreground or background. The algorithm is relatively robust to gradual changes as well as abrupt changes.

2.2.2. Foreground Blob Detection

The goal of this stage is to separate “noise blobs” from relevant objects that are to be tracked. It uses as input the foreground segmentation performed in the background modeling stage. From the foreground segmentation mask, a connected component operation is performed to merge neighboring blobs and a size filter is used to remove small components [13]. From frame to frame, blobs are tracked according to a spatial overlapping rule. If a blob is tracked successfully across a given number of frames, it is added to the tracked blob list, and it is then passed into the advanced tracking stage, described in the next session.



(a) Breaking out (Blob Tracked). (b) Breaking out (Background illustration).

Figure 3. Example of tracking.

2.2.3. Blob Tracking

In the actual blob tracking stage a Kalman filter is used to predict the position of the blob in the next frame. If there is no predicted overlap among any of the blobs in the frame, the tracking is based simply on the connected-component analysis. The position assumed for the person is the bottom of the blob. This position ideally corresponds to the feet of the person, although noise and shadows can cause some shift to it. In practice, this shift is often small and does not significantly affect the overall system performance.

Once a blob is tracked, its pixel coordinates must be transformed to the coordinate frame of the room. For this task, we determine the homography matrix $\mathbf{H}$ describing the projection from the camera plane to the ground plane [14]. An illustration of a tracked blob in a safe-room is given in Figure 3.

2.3. Blob Analysis

For the activity classification, apart from the blob position, we employ the boundary roughness and the area change of the blob over time. For a given fast-moving blob, a change in the area size occurs from frame to frame. Coherent and slow-moving blobs have a less random change in the area size. The normalized area change ΔA_i for the i-th frame is given by

$$\Delta A_i = \frac{|A_i - A_{i-1}|}{A_i} \tag{1}$$

where A_i corresponds to the area of the blob.

For the boundary roughness, we define it as the ratio between perimeter and convex hull perimeter [15]. The convex hull of a set of pixels S is the smallest convex set containing S. The boundary roughness is written as:

$$B_R = P_S / P_{CH_S}$$

where P_S is the perimeter of S and P_{CH_S} is the perimeter of the convex hull of S. To compute the perimeter, a simple approach is to count the number of pixels connected horizontally and vertically plus $\sqrt{2}$ times the number of pixels connected diagonally [15].

Table 1. Confusion matrix for the different activities tested.

		Classified As					
		Entering	Leaving	Fighting	Sleeping	Talking	Breaking Out
Actual	Entering	6	0	0	0	0	0
	Leaving	0	6	0	0	0	0
	Fighting	0	0	4	0	0	1
	Sleeping	0	0	0	5	1	0
	Talking	0	0	0	0	4	0
	Breaking Out	0	2	1	0	0	3

2.4. Classification

In the current implementation, we are using a combination of hard decision rules for the features described in Sections 2.1 and 2.3. Therefore, a set of hierarchical 'IF' conditions is applied in each frame, indicating the most likely action among the ones discussed in Section 3. An obvious extension which can significantly improve the system performance is the use of machine learning methods to combine the features, such as Bayesian classifiers, HMM or support vector machines.

3. Experiments

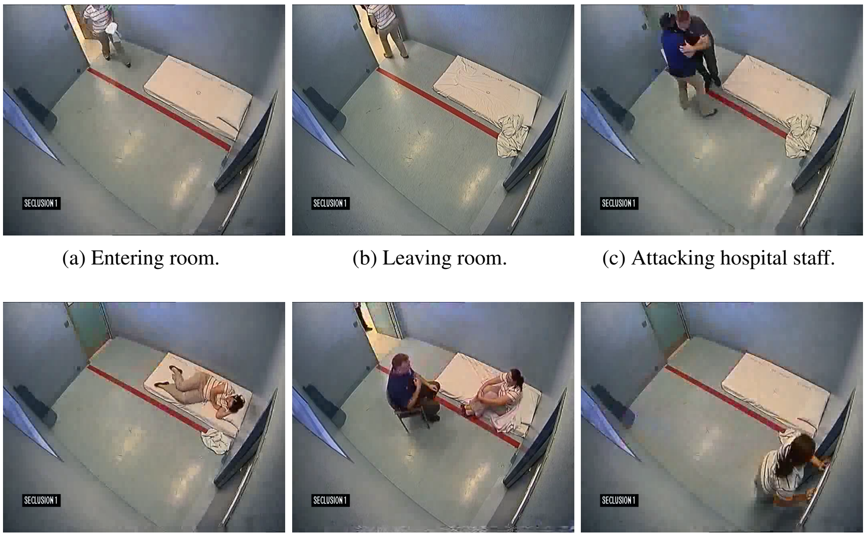
We recorded several short videos at the Emergency Department of the Gold Coast Hospital, in Queensland, Australia. The rooms used were two safe-rooms, where psychiatric patients or patients under the influence of drugs are temporarily kept while in treatment. Due to patient privacy issues, we used doctors and security staff to simulate the actual patients and the different situations, as illustrated in Figure 4. The simulated situations involve: (i) entering the room, (ii) leaving the room, (iii) fighting, (iv) sleeping, (v) talking to staff, and (vi) attempting to breaking out.

For the optical flow analysis, the higher spread in the angle histogram was consistent for all hyper-activity situations, in addition to high flow vector magnitudes. This was combined with the tracked blob analysis, which also assisted in estimating the position of the patient in the room.

For the data set considered, all the activities considered were successfully detected and classified approximately 85% of the time. Most of the classification errors correspond to discriminating between the hyper-activity actions, such as fighting and breaking out attempts. Table 1 shows the confusion matrix obtained for the results in this dataset.

4. Conclusions

We have presented a study on the use of computer vision for monitoring patients in safe-rooms in hospitals. We combined statistics of optical flow vectors with blob tracking to detect activities such as entering and leaving the room, fighting, sleeping, talking, and breaking out attempts. Preliminary results indicated that the system can be potentially applied in a real hospital scenario, helping to prevent injuries to patients and to staff. For a practically low error rate to be achieved, a more sophisticated classification framework



(a) Entering room. (b) Leaving room. (c) Attacking hospital staff.

(d) Sleeping or resting. (e) Talking to staff. (f) Attempt to break out of the room.

Figure 4. Example of activities to be detected.

for combining the features could be employed to improve the system performance. In this sense, we believe that techniques such support vector machines or HMM can be directly added as a module to the proposed approach. Apart from the classification aspect, further work also includes acquiring a more comprehensive dataset such that a more thorough statistical evaluation of results can be achieved.

References

[1] W. Li, Z. Zhang, and Z. Liu, “Expandable data-driven graphical modeling of human actions based on salient postures,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 11, pp. 1499–1510, 2008.

[2] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, “Machine recognition of human activities: A survey,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 11, pp. 1473–1488, 2008.

[3] T. Moeslund, A. Hilton, and V. Kruger, “A survey of advances in vision-based human motion capture and analysis,” *Computer vision and image understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.

[4] M. Enzweiler and D. M. Gavrila, “Monocular pedestrian detecion: Survey and experiments,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, December 2009.

[5] A. Bobick and J. Davis, “The recognition of human movement using temporal templates,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 3, pp. 257–267, 2002.

[6] M. Hu, “Visual pattern recognition by moment invariants,” *Information Theory, IRE Transactions on*, vol. 8, no. 2, pp. 179–187, 2002.

[7] D. Chen, S. Shih, and H. Liao, “Human Action Recognition Using 2-D Spatio-Temporal Templates,” in *Multimedia and Expo, 2007 IEEE International Conference on*. IEEE, 2007, pp. 667–670.

[8] J. Davis and A. Tyagi, “Minimal-latency human action recognition using reliable-inference,” *Image and Vision Computing*, vol. 24, no. 5, pp. 455–472, 2006.

[9] F. Nater, H. Grabner, and L. Van Gool, "Exploiting simple hierarchies for unsupervised human behavior analysis," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2014–2021.

[10] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 9–14.

[11] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey vision conference*, vol. 15. Manchester, UK, 1988, p. 50.

[12] L. Li, W. Huang, I. Y. Gu, and Q. Tian, "Foreground object detection from videos containing complex background," in *ACM International Conference on Multimedia*, 2003, pp. 2–10.

[13] A. Senior, A. Hampapur, Y. L. Tian, L. Brown, S. Pankanti, and R. Bolle, "Appearance models for occlusion handling," *Image and Vision Computing*, vol. 24, no. 11, pp. 1233–1243, November 2006.

[14] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 1st ed. Cambridge University Press, 2004.

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