Automated video surveillance for preventing suicide attempts

Wassim Bouachir
LICEF research center, TÉLUQ
Montréal (QC), Canada
wassim.bouachir@teluq.ca

Rita Noumeir
École de technologie supéieure
Montréal (QC), Canada
Rita.Noumeir@etsmtl.ca

Abstract

Inmate suicide by hanging is documented as a major cause of death in prisons. Important efforts have been made to develop technological prevention tools, but the proposed solutions are mostly using cumbersome devices, in addition to their lack of generalizability. Nowadays, computer vision methods for real-time video analysis have experienced impressive progress. The recent emergence of RGB-D cameras clearly illustrates the achieved advances by offering new ways for machines to interpret human activity. There was however no significant works on exploiting this evolution, and as a result, CCTV systems used for monitoring suicidal inmates are still greatly depending on human attention and intervention. This paper proposes an intelligent video surveillance system for automated detection of suicide attempts by hanging. The proposed algorithm is able to efficiently model suicidal behavior by exploiting the depth information captured by an RGB-D camera. Activity detection is then performed by classifying visual features characterizing body joint movements. Our method demonstrated a high robustness on a challenging dataset including video sequences where suicide attempts are simulated.

1. Introduction

Suicide is known as a common cause of death in correctional facilities and psychiatric hospitals. In Canadian federal prisons, it is the leading cause of un-natural death [2,5]. In most cases, inmates commit suicide by hanging using bedding or clothing, and attempts tend to occur during the night when the number of staff in the service is reduced [15].

Significant efforts have been made to establish prevention techniques, such as CCTV systems used by security staff to monitor persons identified as suicidal. However, these methods depend mainly on human analysis and intervention, and nearly one third of the attempts continue to occur in areas with an enhanced level of observation and monitoring [5], which calls into question the effectiveness of conventional prevention methods. Several other devices have been designed to introduce automation in detecting hanging attempts, such as special articles of clothing to be worn by inmates [13], a door alarm system to be activated if the door is closed with something (e.g. bed sheet) over its top [4], and a bracelet for monitoring the person's oxygen saturation level and pulse rate [1]. Such systems either require wearing bulky equipment or they are designed for very specific situations. To the best of our knowledge, the only work that addressed suicide detection in an automated vision-based approach is that of Lee et al. [10], where they propose to analyse 3D images captured by the Asus Xtion camera. However, this is a preliminary and limited study considering only the case of partial suspension, and not addressing real-world settings difficulties such as viewpoint and scale change. Moreover, their method does not take into account realistic scenarios, as each video sequence is captured for a fixed duration of 3 seconds.

Our paper presents a new intelligent surveillance system for detecting suicide attempts by hanging. The system is designed to operate autonomously in real-world indoor conditions, in order to activate an alarm if the event of interest is detected. Currently, most commercial video surveillance systems use visible-light cameras and basic image processing methods to recognize simple events (e.g. intrusion). Understanding a complex activity is a very challenging task due to the difficulty of analyzing a 3D scene projected on bi-dimensional images. Furthermore, 2 other significant challenges are considered in this work:

- The event of interest should be detected within a short period of time, before the person completes his action;

- The system should be able to operate under various illumination conditions, especially in the darkness.

The rest of this paper is organized as follows. In the next section, we review related works on vision-based activity recognition. The proposed method for suicide detection is presented in section 3. Experimental results are provided in section 4, and section 5 concludes the paper.

2. Related works

Behavior and event analysis is an active research topic within the computer vision community. This interest has

been driven by the potential for many applications, and especially intelligent surveillance. Generally, an event recognition system operates on three steps corresponding to its major components: 1) image analysis for visual feature extraction, 2) feature processing for behavior modeling, and 3) event recognition that is mostly achieved through classification algorithms or exemplar matching methods. According to the type of features extracted, existing works on behavior recognition can be divided into three categories.

The first category uses global models to represent the human behavior by the appearance and movements of the whole human body region, without identifying its different parts. Activities are thus represented by global templates where various appearance and/or motion features can be computed. For example, the authors in [6] compute optical flow features to be compared against templates stored in an exemplar database. Another representative work in this category is that of Blank et al. [7], where actions and activities are represented as space-time shapes. Since most algorithms in this category depend on background subtraction, the proposed methods lack the flexibility to handle challenging cases such as dynamic backgrounds, camera motion and silhouette deformation. Moreover, global appearance models are not suited for partial occlusion, which may be a frequent situation in realistic settings.

Unlike global methods, representations of the second category use local features that provide a better tolerance to certain conditions, such as illumination changes, deformation, and occlusion. The corresponding recognition systems rely on low-level and mid-level features like space-time interest points [9] extending the notion of spatial interest points into the spatio-temporal domain, and dense points [14] that are sampled and tracked on subsequent frames. Despite encouraging results obtained in realistic scenarios, these methods often fail to analyze complex human behavior because of the limited semantics they represent [11].

The works of the third category use pose estimation methods to represent activities by employing human body part structure and positions. This representation is an alternative to local feature models, and it conforms to the early Johanssons studies on how human understand actions [8]. In [8], the author demonstrated that humans can easily recognize events from the motion of a few human body joints. Starting from this ascertainment, numerous activity recognition algorithms used pose estimation techniques to locate human body joints. Among the most significant achievements, we mention the method of Yilmaz et al. [16], and that of Cherian et al. [3]. Both works rely on the detection of human body joints. These methods were tested offline on several datasets, showing a good recognition performance. However, they highly depend on pose estimation to locate human body joints. The detection of body joints in itself is an open computer vision problem, requiring a significant computation time and involving very challenging techniques, such as image segmentation, human tracking, and 3D reconstruction, which are still open and active research areas.

In our work, we are particularly interested to methods and techniques of the latter category, as pose based-approaches are more appropriate to model complex real-life activities. One of the major issues with using human body joints is that pose estimation on color images is a computationally expensive task. For instance, the method of Cherian et al. [3] requires about 3 seconds estimating a pose model on a single video frame, which is inappropriate for real-time surveillance applications. Our idea to tackle this challenge consists in exploiting depth maps that can be obtained using recent cost effective RGB-D cameras. Such acquisition devices provide the 3D spatial information on human body, in addition to the color image of the scene. We thus avoid analyzing the projection of a 3D world on bi-dimensional images, which is the source of the computational complexity problem. Moreover, with RGB-D cameras, pose estimation can be achieved in real-time under various illumination conditions. Detailed methodological aspects are discussed in the next section.

3. Proposed method

Recently, RGB-D cameras provided new opportunities to address the pose estimation problem in real-time. These advances offer novel possibilities for dealing with vision-based activity recognition. Our method for activity modeling is based on the exploitation of the 3D visual content captured with a low-cost RGB-D camera. More specifically, we use human joint relative positions in the 3D space to compute pose and motion features during movement, in order to learn the activity of interest. To detect suicidal behavior, our recognition algorithm processes each current observation to perform a binary classification. The activity of interest is finally detected if the percentage of positive observations exceeds a certain threshold during a sliding temporal window.

3.1. Pose estimation

To construct our visual representation, we consider the human body as an articulated structure of segments connected by joints. Our activity of interest is thus represented by the evolution of the joints spatial configuration during a time interval. Shotton et al. [12] demonstrated that 3D joint localization can be achieved in real-time by using RGB-D imaging. In their algorithm, body part recognition is firstly performed with pixel-level classification from single depth images. The obtained result is considered as a transitional representation for the 3D pose. A mean shift mode detection is then applied to find local centroids of the body part probability mass and generate hypothesis for the 3D locations of body joints. Finally, a skeleton model is fitted to the 3D joint positions proposals, by taking into account the temporal information from previous frames and kinematic constraints. In the proposed system, this algorithm is used as the first stage of the pipeline. An important benefit of this method is that only the last step (i.e. fitting the skeleton model) uses the temporal information, while both body

part labels and 3D joint proposals are computed from single depth images processed separately, which results in enhancing the system's ability to recover from errors. For visual representation, we consider a subset of $N = 16$ upper body joints. For each frame t, the 3D joint coordinates are extracted as:

$$X_t = \{J_t^i = (x_i, y_i, z_i)_t | i = 1, 2, ..., N\}. \qquad (1)$$

These positions are used to compute feature vectors.

3.2. Pose and motion features

Once joint positions are extracted, we compute their pairwise relative positions by calculating the 3D distances corresponding to pairs of joints. The obtained features provide a discriminative representation for the suicide activity. As an illustration, the activity of interest includes the action of placing a strangling object around the neck, which could be perceived as bringing the hands to the neck from top to bottom.

Concretely, we use the joint positions to compute 2 types of feature vector:

- $P_t = \{dist(J_t^i, J_t^j) | i, j = 1, 2, ..., N; i \neq j\}$ that describes the pose at frame t using the 3D distances $dist(J_t^i, J_t^j)$ between pairs of joints J^i and J^j on the current frame t.

- $M_t = \{dist(J_t^i, J_{t-1}^j) | i, j = 1, 2, ..., N\}$ that describes the motion performed between 2 subsequent frames using the 3D distances between joints on the current frame t and those on the previous frame $t - 1$.

The feature vector P_t captures the pose property in the current frame t while M_t captures the motion property. We combine the two feature subsets P and M, having respectively $C_N^2 = 120$ and $N^2 = 256$ elements, in a single 376-dimensional vector to represent each frame as $F_t = [P_t, M_t]$. We note that the proposed features are invariant against orientation, since we perform pair-wise comparisons instead of directly using joint positions (in the 3D camera coordinate system). We also normalize the computed distances with respect to the person's height to ensure scale invariance.

3.3. Feature learning and classification

In our conception, recognizing the activity of interest involves classifying single observations F_t at time t as one of the 2 classes $'suicide'$ and $'unsuspected'$. For this purpose, we construct a Linear Discriminant Analysis (LDA) classifier by using the computed features F_t as the classification model variables. The feature set of each class is thus modeled as a multivariate normal distribution with a common covariance matrix and 2 different mean vectors. These parameters are estimated from labeled data during the training step. The entire offline training procedure is presented in figure 1.
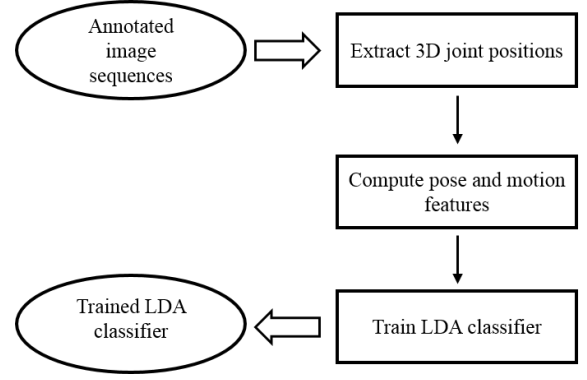

Figure 1: The offline training procedure. Joint positions are extracted in the 3D space in order to calculate feature vectors encoding the joint relative positions. The dataset is then used to learn an LDA classifier.

During real-time detection, new observations corresponding to single frames are classified as the class having the nearest mean vector according to Mahalanobis distance. Due to its low computational complexity, this classification procedure meets real-time processing requirements, while its performance is comparable to more sophisticated algorithms that we tested.

3.4. Activity detection

Given a trained LDA classifier, the main recognition algorithm processes a depth image stream in an on-line manner to detect the activity of interest. First, the body pose is estimated for each retrieved depth image by [12]. The 3D joint positions are then used to compute the feature vector F_t considered as the current local observation. Local observations are processed sequentially to be classified using LDA as positive or negative observations. Suicide detection is based on observing the person's activity during a sliding temporal window of width Δ_o. At each iteration, the window is shifted by the temporal step S. Finally, a suicide attempt is detected if the percentage of positive observations θ_t exceeds a threshold θ_{min}. The main on-line detection algorithm is summarized in figure 2. More details on the processing steps are provided in Alg. 1.

4. Experiments

4.1. Dataset

Since there is almost no work in the computer vision literature on suicide detection, there are no publicly available datasets that we can use to evaluate such an activity recognition algorithm. We thus created our dataset in a room where dimensions are close to those of a prison cell. Our RGB-D sensor is the Kinect v2 camera that we placed in a corner near the ceiling, at a distance of approximately 0.30 m from the ceiling, with a tilt angle of $35°$. This setting allows capturing a person at a maximum distance of 4.5 m. For efficient body pose estimation, the distance between the moni-
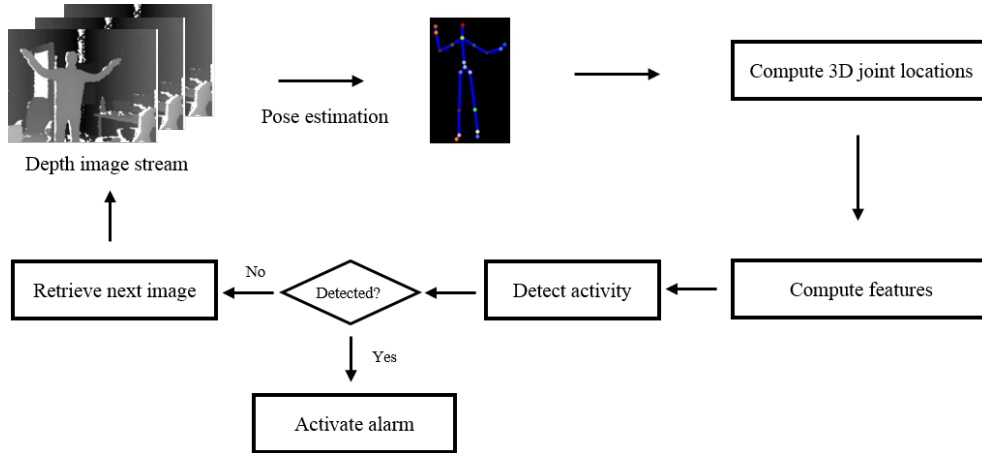
Figure 2: Real-time activity detection algorithm. Pose estimation is performed in real-time using depth maps. Once feature vectors encoding the joint relative positions are computed, activity detection is based on classifying local observations subsequently.

Algorithm 1 On-line activity detection

Input: depth frame t
Output: detection result
Assumption: processing frame t with $t >= 1$
Initialization: detected = false; $\theta_t = 0$
1: **while** $detected == false$ **do**
2: - Estimate pose
3: - Compute joint locations
4: - Calculate pose feature vector P_t
5: - Calculate motion feature vector M_t
6: - $F_t = [P_t, M_t]$
7: - Classify F_t using LDA
8: - Update θ_t
9: **if** $\theta_t \geq \theta_{min}$ **then**
10:    - $detected = true$
11: **else**
12:    - Shift temporal window by S
13:    - Retrieve frame $t + 1$
14: **end if**
15: **end while**

---

tored person and the camera should be between 0.5 and 4.5. Moreover, we experimentally verified that optimal joint localisation is obtained at a distance of 1-4 m from the Kinect. A video dataset was created with the participation of 21 persons. Each participant was asked to perform two scenarios.

- In the first scenario, the participant shows an unsuspected behaviour, such as moving in the room, sitting, removing a piece of clothing, wearing clothes, etc. During these actions, the actor may have different orientations with respect to the camera;

- For the second scenario, the participant simulates a suicide attempt by hanging in three steps: 1) he tries to create a hangman's knot using a bed sheet (each participant could create the knot in a different way), 2) he

attaches the knot to a fixed point in the room, and finally 3) the knot is placed around the neck.

Note that for training the system, we consider learning the activity of attaching the strangling object to a fixed structure and placing it around the neck. Our dataset including 42 video sequences is used to train the system and evaluate the recognition algorithm.

4.2. Results

We randomly divided the video dataset into two disjoint subsets: 32 video sequences are used for training the system, while 10 sequences are used exclusively for testing. We implemented our activity recognition algorithm using Matlab installed on a 3.6 GHz Core i74790 CPU where processing is done.

We applied the learning procedure presented in figure 1 on the training set to extract the feature vectors corresponding to the sequence frames. By using the training set, we learned LDA classifier and performed tests on the remaining 10 test sequences. Note that LDA was selected among several simple classifiers based on its performance in classifying single frames as one of the 2 classes $'suicide'$ and $'unsuspected'$. The sliding temporal window Δ_o in the main recognition algorithm was set to 5 seconds. At each iteration, the temporal window is shifted by a step $S = 0.07$ second. Finally, a suicide attempt is detected if the current indicator θ_t exceeds $\theta_{min} = 0.7$ for the current position of the sliding window. Table 1 presents detection results for 5 video sequences where participants simulate suicide attempts using a bed sheet. The 5 suicide attempts were simulated by five different participants. We used the same camera settings for all video sequences, but the participants were asked to proceed in different ways to the construction of the hangman's knot, with changing body orientation (which determines the camera view angle), and different fixation points where the knot is attached. Figure 3 shows examples

Video	Duration (s)		Detection	
	Start	End	Yes/No	Time (s)
1	50	68	Yes	52
2	21	34	Yes	27
3	22	32	Yes	25
4	82	92	Yes	91
5	13	24	Yes	16

Table 1: Recognition results for test sequences where suicide is simulated. The suspected activity includes attaching the knot and placing it around the neck. For each video, the table shows: the start time of the suspected activity (Start), end time (End), detection result (Yes/No), and the detection time (Time). Times are expressed in seconds.

of depth frames from a suicide sequence and corresponding body skeletons.

As shown in table 1, the activity of interest was recognized in all scenarios. Detection was achieved mostly during the first few seconds. In scenario 4, the relatively delayed detection is mainly caused by the person's orientation with respect to the camera view angle. In such a case, the side viewpoint caused the occlusion of an upper limb, which resulted in affecting the joint localization accuracy. The accuracy in estimating the joint positions has also been slightly affected in sequence 2, despite a front view. This situation can be explained by the difficulty in distinguishing the body from the background when their depth difference is minimal (i.e. the body is very close to the background), which represents a common limitation for time-of-flight cameras. Note that even relatively delayed detection (scenarios 2 and 4) is considered as early enough to activate an alarm and thus allow reactive intervention. Indeed, the major cause of death following a suicide attempt in the considered environment (e.g. psychiatric hospital room, prison cell) is the occlusion of blood vessels and/or airway, which takes a few minutes once the knot is sufficiently tightened around the neck. This process generally takes relatively longer duration compared to the case of death by cervical fracture, which generally requires releasing the body from a high position.

Table 2 presents detection results for the 5 remaining normal behavior sequences. With these videos, we noticed that the main difficulty for the recognition algorithm is to avoid confusing between suicidal behavior and other activities of daily living requiring to move the hands around the neck. We therefore asked participants to wear or remove pieces of clothing during unsuspected scenarios. The recognition results include a single false alarm where such actions were confused with suicidal behavior, while detection results were correct for 4 sequences.

Statistical measures summarizing the recognition tests are provided in table 3. The overall accuracy of the recognition algorithm is 90%. The proposed system also demon-

Video	False alarm	Time (s)
6	No	-
7	No	-
8	No	-
9	Yes	43
10	No	-

Table 2: Recognition results for test sequences showing normal behavior (unsuspected). For each video sequence, the table shows the detection result. In the case of false detection, the detection time is indicated in seconds.

TPR	TNR	FPR	FNR	ACC
1	0.8	0.2	0	0.9

Table 3: Statistical measures of recognition performance for 10 test sequences. TPR: True Positive Rate (sensitivity), TNR: True Negative Rate (specificity), FPR: False Positive Rate, FNR: False Negative Rate, ACC: Accuracy (correct detection rate).

strated high sensitivity with a True Positive Rate (TPR) of 100%, while the percentage of unsuspected sequences that are correctly recognized as such (specificity) is 80%. It should be noted here that the system sensitivity is a very important indicator since ideally all suicide attempts should be detected, while a certain percentage of false alarm can be tolerated.

5. Conclusion

We proposed a novel method for detecting suicide attempts based on body joint movements. Joint positions are extracted in real-time from depth images captured by the Microsoft Kinect camera. Once the system is trained to recognize visual features corresponding to suicide actions, on-line detection is performed through a simple yet efficient recognition algorithm. We obtained a high accuracy and a 100% sensitivity on a challenging dataset with considerable variations between simulated suicide scenarios.

Our future work will focus on reinforcing the proposed algorithm in order to maximize the detection chance in a short observation time. An interesting avenue is to model the interaction between the human body and the strangling object before attaching it (i.e. when the knot is being created). This can be achieved by combining depth information with corresponding infrared images provided by the kinect, on which the knot could be detected.
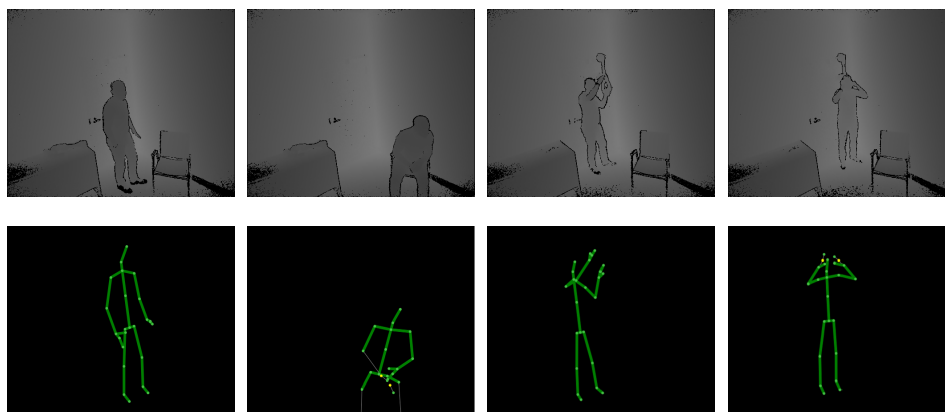
Figure 3: Examples of depth frames from a suicide sequence (first row) and corresponding body skeletons where joints are located (second row).

Acknowledgement

References

[1] M. Bailey. Method of preventing an inmate from committing suicide, Oct. 27 2011. US Patent App. 12/799,243.

[2] B. E. Burtch and R. V. Ericson. *The Silent System: An Inquiry Into Prisoners who Suicide and [an] Annotated Bibliography*. Centre of Criminology, University of Toronto, 1979.

[3] A. Cherian, J. Mairal, K. Alahari, and C. Schmid. Mixing body-part sequences for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2353–2360, 2014.

[4] F. Cook. Door suicide alarm, Mar. 5 2013. US Patent RE44,039.

[5] Correctional investigator Canada. A three year review of federal inmate suicides (2011-2014), 2014.

[6] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, pages 726–733. IEEE, 2003.

[7] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE transactions on pattern analysis and machine intelligence*, 29(12):2247–2253, 2007.

[8] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14(2):201–211, 1973.

[9] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[10] S. Lee, H. Kim, S. Lee, Y. Kim, D. Lee, J. Ju, and H. Myung. Detection of a suicide by hanging based on a 3-d image analysis. *IEEE sensors journal*, 14(9):2934–2935, 2014.

[11] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1234–1241. IEEE, 2012.

[12] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1297–1304, June 2011.

[13] A. Tasezen and R. E. Schilling. Suicide prevention clothing, Feb. 19 2013. US Patent 8,375,466.

[14] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3169–3176. IEEE, 2011.

[15] World Health Organization and the International Association for Suicide Prevention. Preventing suicide in jails and prisons, 2007.

[16] A. Yilmaz and M. Shah. Recognizing human actions in videos acquired by uncalibrated moving cameras. In *Tenth IEEE International Conference on Computer Vision (ICCV'05)*, volume 1, pages 150–157. IEEE, 2005.