Intelligent video surveillance for real-time detection of suicide attempts

Wassim Bouachir[a,*], Rafik Gouiaa[b], Bo Li[b], Rita Noumeir[b]

[a]*LICEF research center, TÉLUQ University Montréal (QC), Canada*
[b]*École de technologie supérieure, Montréal (QC), Canada*

ABSTRACT

Suicide by hanging is a sentinel event and a major cause of death in prisons, with an increasing frequency over recent years. The rapid detection of suicidal behavior can reduce the mortality rate and increase the odds of survival for the suicide victim. Significant efforts have been made to develop technologies for preventing hanging attempts, but most of them use cumbersome devices, or they are mainly depending on human attention and intervention. In this paper, we propose a vision-based method to automatically detect suicide by hanging. Our intelligent video surveillance system operates using depth streams provided by an RGB-D camera, regardless of illumination conditions. The proposed algorithm is based on the exploitation of the body joints'positions to model suicidal behavior. Both dynamic and static pose characteristics are calculated in order to efficiently capture the body joints'movement and model usicidal behavior. Results from the experiments on realistic video sequences, show that our system achieves a high accuracy in detecting suicide attempts, while meeting real-time requirements.

1. INTRODUCTION

Suicide is a major cause of premature and preventable death within the correctional settings. For instance, inmate suicide was reported as the leading cause of death in Canadian prisons between 1994 and 2014 (of the Correctional Investigator of Canada, 2014). In addition, about one-third of prison deaths were caused by suicide in the USA between 2000 and 2010 (Noonan and Ginder, 2013). Generally, hanging is the most common method of inmate suicide, where inmates use clothes, cord or bed linen to hang themselves. Suicide attempts tend to occur when the victims are being held in isolation or segregation cells, and during the night or weekend when the security staff is the lowest (World Health Organization).

To deal with the phenomenon of suicide by hanging in prisons, correctional officers tend to establish, in the first place, a suicide-safe environment. This can be a cell where hanging points and unsupervised access to lethal materials are eliminated or minimized. As the technology is developing rapidly, video surveillance systems (e.g CCTV) have been used, for a long period, as an alternative to the direct observation by correctional staff. However, camera blind spot together with busy camera operators can seriously limit

[*] LICEF research center, TÉLUQ University Montréal (QC), Canada
e-mail: wassim.bouachir@teluq.ca (Wassim Bouachir)

the performance of such systems. Tragically, numerous examples of suicides continue to occur in full view of camera equipment, which brings up the question of the effectiveness of such vision-based systems for preventing suicide by hanging attempts in jails and prisons (of the Correctional Investigator of Canada, 2014). Numerous other solutions have been established for the purpose of detecting and preventing suicide by hanging attempts. For instance, a special protective clothing (Safety smocks and blankets) has been designed to be worn by actively suicidal inmates (Hayes, 2013). A top door alarm (Cook, 2011), which triggers an alarm if the door is used as a ligature point, allowing for a life-saving proactive emergency response. Moreover, the *'bracelet for life'* (World Health Organization, 2007) has been used for monitoring several physiological parameters. If prisoner's vital signs are detected as being outside the normal range, an alarm is triggered and an emergency response is activated. Despite their effectiveness in detecting a number of cases of suicide by hanging, establishing such systems in practice is very challenging. In fact, most of these systems require either wearing cumbersome equipments or they are a source of false alarm even if the inmate simply removes the equipment (bracelet, clothes etc).

With the increasing development of new technologies for human action recognition, a few solutions have been proposed in the aim of improving the existing video-surveillance systems to automatically detect suicidal behaviors and trigger an alarm. In this sense, (Lee et al., 2014) presented a method for automatically analyzing depth images captured by an Asus Xtion Pro camera, and detect suicidal behavior. However, this is a preliminary work, where only the case of a partial suspension hanging was considered, without dealing with real-world difficulties such as occlusion and scale invariance. Besides, only a few video sequences with a short duration of 3 seconds each one, have been used to evaluate the proposed algorithm. Lately, we presented an intelligent video-based system for automated detection of suicide by hanging attempts (Bouachir and Noumeir, 2016). Unlike in (Lee et al., 2014), we performed our experiments on a large dataset captured by an RGB-D camera, where 21 persons are asked to simulate different scenarios for unsuspected behavior and suicide attempts. Furthermore, video sequences have been captured, where challenging real-world conditions such as occlusions, illumination changes, and scale changes have been considered. Our system is also able to operate day and night without bothering the prisoners, owing to the invisible illumination used by the Kinect camera. Despite the promising results given by our algorithm, it faced difficulties for recognizing some daily living activities such as wearing or removing clothes, which involve movements similar to the movement of the body during suicide by hanging attempts. In this paper, we aim to extend our previous work in the following directions:

- A real-time scaling algorithm is proposed to deal with the effect of morphological difference within candidates and keep the features invariant with respect to people.

- A feature selection approach is applied in order to speed up our algorithm for real-time application and improve its generalization capacity.

- We analyze the performance of numerous classifiers for detecting suicide attempts presented in realistic videos.
- We propose to oversample the minority class to overcome the problem of the unbalanced dataset.
- We perform a more complete performance evaluation using 2 strategies: a frame-based evaluation and a sequence-based evaluation.

2. Related works

In the past few decades, human action recognition has drawn much importance in the field of video analysis technology owing to the increasing demand from a wide range of applications (e.g human computer interaction, ambient assisted living and video surveillance). In the context of video surveillance applications, the automatic detection of abnormal behaviors can be used to alert the security staff of a potential danger such as reporting an inmate committing a suicide by hanging. Generally, a camera-based action recognition system operates on two main steps: 1) feature extraction which consists in building visual cues intended to be informative and non-redundant with respect to the corresponding activities, and 2) action learning and classification which based on training classifiers using the extracted features, and using them for classifying observation. According to the type of the extracted features: human action recognition methods can be categorized into 3 groups (Weinland et al., 2011) as follows: The first group used global approaches that consist on detecting the whole human body without caring about identifying and labeling the individual body parts. Numerous type of features have been used such as silhouette (Gouiaa and Meunier, 2015), contour (Cheema et al., 2011) or optical flow (Fathi and Mori, 2008). Most proposed methods in this category operate on the whole human body and obey to a global representation, which is not flexible enough to capture intra-class variations of activities.

Unlike the first category, methods in the second category mainly focus on designing a local representation of human actions instead, where the image/video are divided into small patches, regardless the body parts labeling, illumination changes or the body localization. These small patches catch the regions of high variations in time and spatial domain and involve appearance and/or motion information. For instance, the space-time interest descriptors (Yang and Tian, 2014) were proposed to generalize the concept of interest points and local descriptors (Bay et al., 2006). Despite their effectiveness to overcome some global representation limitations, including noise and partial occlusion, such methods still only contain limited spatiotemporal information, which is insufficient for representing complex activities.

The methods in the third group are based on pose estimation approaches to represent a human action as a sequence of poses over time. This modeling is often done by exploiting the spatial configuration of human body structure. This representation is derived from the principle, published in (Johansson, 1973), explaining how humans observe actions. This work demonstrates that humans are able to identify an action from a few poses only. Several methods have been proposed in this context to tackle the problem of human action recognition using human body joints localization. For instance, (Parisi and Wermter, 2013; Parisi et al., 2015) proposed methods for human action recognition based on the clustering of pose and motion features with self-organized maps. The reader can find further details in the survey (Sarafianos et al., 2016). However, pose-based human action recognition can be difficult due to the complexity of getting high quality poses from RGB videos in a real-world scenario, except in controlled environments (e.g static and calibrated cameras and simple backgrounds). Pose-based activity recognition is still a challenging research area, demanding significant improvements to deal with numerous limitations.

In this paper, we are interested in methods of the third category, since they are more suitable to represent complex human behaviors. To deal with the limitations of such systems, we propose to use images captured by a depth camera, providing the 3D spatial information in addition to the RGB images. The proposed method is detailed in the next section.

3. Proposed Method

The relevant cues provided by the depth information to localize the body joints can help dealing with some of the previously described limitations and open a new line of work to tackle the human activity recognition problem. The proposed method relies on using the spatial configuration of body joints in the 3D space to compute pose and motion features. To detect suicidal behavior, the extracted features corresponding to the current observation are fed to a binary classifier. A suicide by hanging attempt is detected if the percentage of positive observations exceeds a certain threshold during a sliding temporal window.

3.1. Pose representation and analysis

The human body is modeled as an articulated structure of rigid segments connected by joints. Thus, the action of interest can be seen as the temporal evolution of the joints'spatial configuration. With the emergence of depth cameras, the provided 3D depth data largely facilitates the task of human body parsing. Moreover, it has helped to develop a rather powerful human motion capture technique (Shotton et al., 2013) that outputs, in real time, the 3D joints'positions of the human skeleton. In our algorithm, this method is applied on the depth images to obtain the 3D location of joints'human body.

We experimentally verified that the lower body joints are not relevant for our joint pairwise distance features. Indeed, the pairwise distances between the lower body joints remain almost constant, as the spatial configuration of the lower body parts does not undergo any significant changes during the activity of interest. We therefore propose to track only the upper body parts. Thus, we consider a subset of $N = 16$ upper body joints (See Fig.1). In each frame t, the 3D joint coordinates are noted as:

$$X_t = \{J_t^i = (x_i, y_i, z_i) | i = 1 \cdots N\} \tag{1}$$

Where (x_i, y_i, z_i) are the 3D coordinates of the *ith* joint J at time t which is noted as J_t^i.

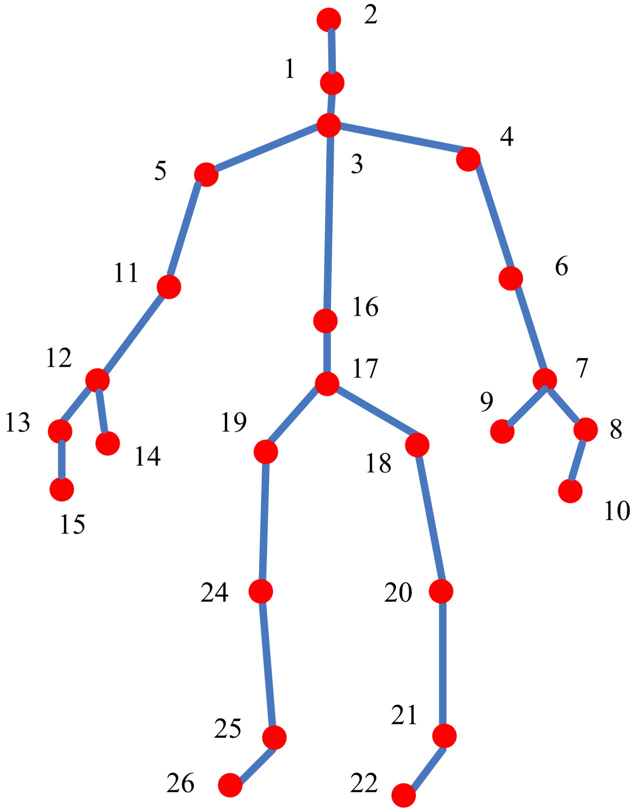

Fig. 1: Location of human joints given by the method of Shotton et al. (Shotton et al., 2013).

3.2. Pose and motion features

Usually, inmates commit a suicide by hanging by placing a strangling object around the neck (World Health Organization, 2007). This typically requires moving the hands to the neck from top to bottom, implying the movement of several upper body joints. Based on this, we model the upper body configuration change using the relative distances between joints. The 3D joints are used to calculate two different features:

- $P_t = \{dist(J_t^i, J_t^j)|i, j = 1 \cdots N; i \neq j\}$, which is pairwise disjoint distances between the list of joints, used to describe the pose at frame t.

- $M_t = \{dist(J_{t-1}^i, J_t^j)|i, j = 1 \cdots N\}$, which is pairwise distances between the list of joints in frame t and frame (t-1), used to describe the elementary motion performed between 2 subsequent frames.

For each frame t, we combine the pose features vector P_t and the motion feature vector M_t having respectively $C_N^2 = 120$ and $N^2 = 256$ components to form a single 376-dimensional vector $F_t = (P_t, M_t)$. Note that the feature vector F_t is invariant to rotation, since we perform pairwise comparisons instead of directly using joint positions (in the 3D camera coordinate system). To ensure scale invariance, we normalize each feature vector F_t by a scaling parameter, estimated as follows.

3.3. Scaling parameter estimation

Scale invariance is considered as an important criterion for designing robust feature descriptors. In our case, the pairwise distances between the 3D joints within one frame or across a sequence of neighboring frames are largely variant with respect to the monitored person. For example, the distance between shoulder and elbow of people with large body size is greater than that of people with smaller body size. To deal with this difficulty, we propose to normalize the computed feature vector in order to remove the influence of body size. The normalization is achieved by dividing each feature vector by the corresponding distance between the neck and the spine middle. Our choice is explained by 3 reasons:

- The distance between these two joints is proportionate to the person's height,
- For one specific person, the distance between the neck and the spine middle is stable and does not dramatically changed with respect to the current pose,
- The neck and middle spine joints are always visible for a Kinect camera. As shown in Fig.2, we see that three joints, which are head, neck and spine middle are always observable. This gives us a heuristic basis for choosing the distance between the neck and the spine middle as a scaling parameter noted s.

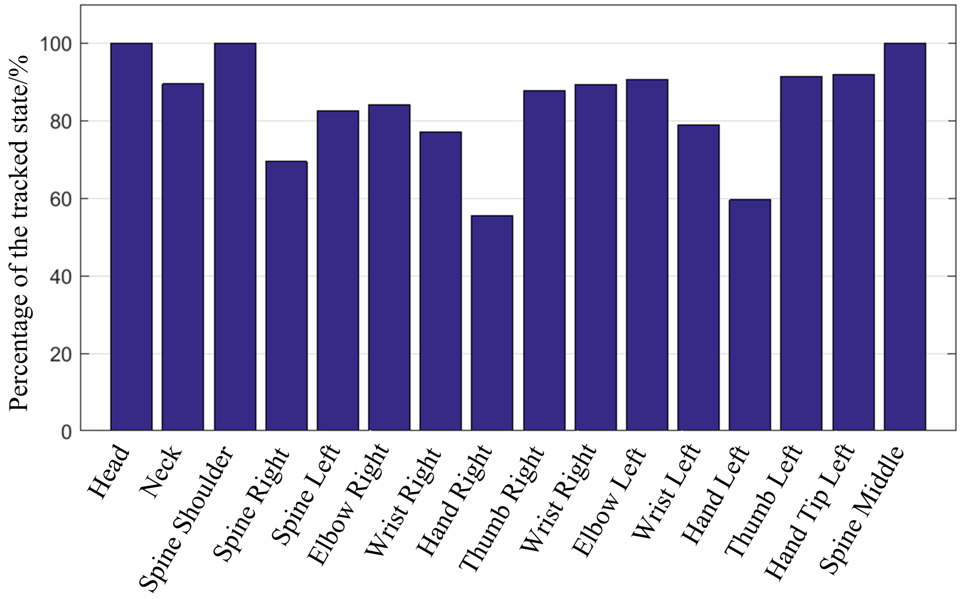

Fig. 2: Percentage of tracked instances for each joint.The total number of frames is 51003.

Note that the scaling parameter can be simply calculated from one frame. However, using a sequence of frames, we propose to estimate s by using the median value of the list of distances generated across the sequence. The main reason for choosing the median value is that it is robust to outliers. For this, we implement an algorithm based on Min-Max heaps for estimating the median value in a constant time (see algorithm 1).

Algorithm 1 Online scaling parameter estimation

Input: neck position and spine middle position from current frame; min heap and max heaps
Output: estimated scaling parameter
Assumption: the size difference of min and max heaps is less than or equal to one. Both min and max heap are not full

```
Calculate the distance $d$ between the neck and spine middle
if the sizes of min heap and max heap are not equal then
    Add $d$ to the heap with the smaller size
    The scaling parameter is the average of the root values from both heaps
else
    Add $d$ to min heap
    The scaling parameter is the value of the root of min heap
end if
if Either min or max heap is with full size then
    Abandon the last half elements in both min and max heap
end if
```

In this algorithm, one min-heap and one max-heap are used to dynamically and adaptively estimate the median value. Both heaps are maintained to make their size difference no larger than one. In this case, the median value is trivially given as follows:

- If both heaps have the same size, the median value is the averaged of the two roots.
- Otherwise, the median value is the root value of the heap with a larger size

In addition, if both heaps are full, we abandon the second half of elements in both heaps to allow the median to be updated from the new measurements. Once the scaling parameter is available, current feature sample is scaled as:

$$SF_t = \frac{F_t}{s}. \tag{2}$$

3.4. Feature selection and learning

Generally, a large number of features leads to the over-fitting problem and consequently decreases the generalization capacity of the machine learning model. To overcome this problem, various dimensionality reduction approaches have been proposed. Feature selection is one of the most popular techniques used for dimensionality reduction, by removing noise and redundant feature. The goal is to select a small subset of features from the original ones according to certain relevance criterion, which leads to higher generalization capacity, lower running time, and better model interpretability. According to the learning approach (i.e supervised, unsupervised or semi-supervised), different feature selection algorithms have been proposed. In the context of supervised learning, feature selection techniques can be categorized into filters, wrappers, and embedded techniques. In this work, we are particularly interested in methods of the first category due to their simplicity and computational efficiency. This method aims to identify a subset of features among a possibly large set of features that are relevant for predicting a response, according to the score attributed to the individual features. However, using only a scoring function to decide the features relevance can lead to a rich redundancy when the features are very dependent. A popular filter technique that is able to handle the redundancy among the selected feature is the

Minimum Redundancy Maximum Relevance (mRMR) (Peng et al., 2005). Using this algorithm, the subset features $\{x_j | j = 1 \cdots m\}$ are sequentially selected as follows:

$$\max_{x_j \in \Lambda \setminus \Gamma} = \left[I(x_j; c) - \frac{1}{|\Gamma|} \sum_{x_i \in \Gamma} I(x_i; x_j) \mid i = 1 \cdots |\Lambda| \right] \tag{3}$$

where $I(x_j; c)$ is the mutual information value between individual feature x_i and class c, $I(x_i; x_j)$ is the mutual information between two features x_i and x_j, Γ is the subset of the best features, Λ is the original set of features with $|\Lambda|$ cardinality.

The idea behind the mRMR is to sequentially select a feature that is relevant with respect to the target class c (max-relevance) and simultaneously has a low dependence to the already selected features in Γ (min-redundancy). Both criterions are evaluated based on the mutual information. In the literature, different methods have been proposed to ensure the max-relevance (correlation between a feature vector and the target class c) criterion such as Fisher score, t-statistics, and entropy score. Thus, in this work, we consider a generalized version of the mRMR algorithm presented by the following equation:

$$\max_{x_j \in \Lambda \setminus \Gamma} = \left[S(x_j) - \frac{1}{|\Gamma|} \sum_{x_i \in \Gamma} I(x_i; x_j) \mid i = 1 \cdots |\Lambda| \right] \tag{4}$$

where $S(x_j)$ is the score of the jth feature.

Feature selection is a crucial step that can help us for speeding up our algorithm to meet the real-time requirements and improve the learning performance of our system, by removing the redundancy. Thus, once the feature selection is performed, we obtain a new observation $\widehat{SF}_t$ at time t, with a smaller dimension ($|\widehat{SF}_t| < 376$).

In our system, activity recognition is achieved by applying a binary classification on a single observation $\widehat{SF}_t$ in order to categorize it as *'suicide'* or *'unsuspected'*. For this purpose, we tested several binary classifiers which are constructed and tuned using the extracted features $\widehat{SF}_t$ of the training set. For this purpose, different binary classifiers including Linear Discriminant Analysis (LDA), Linear Support Vector Machines (L-SVM), Support Vector Machine with radial basis kernel (RBF-SVM) and Naive Bayes (NB) have been constructed and tuned using the extracted features $\widehat{SF}_t$ of the training set. Due to its better performance, we chose RBF-SVM which outperforms the classifiers that we tested. Moreover, we note that due to the real-time requirements, we only tested bit complex classifiers. The flowchart of the offline training procedure is depicted in Fig. 3.

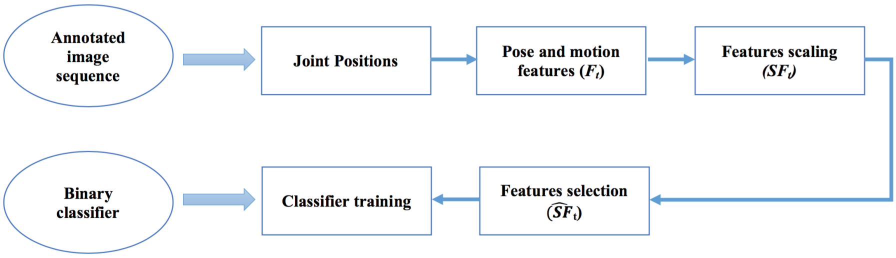
Annotated image sequence
Joint Positions
Pose and motion features (F_t)
Features scaling (SF_t)
Binary classifier
Classifier training
Features selection ($\widehat{SF}_t$)

Fig. 3: A flowchart of the offline training procedure.

Algorithm 2 On-line suicide by hanging detection

```
Input: depth frame $t$
Output: detection result
Assumption: processing frame $t$ with $t >= 1$
Initialization: detected = false; $\theta_t = 0$;

While detected==false do
- Estimate pose
- Compute joint locations
- Calculate pose feature vector $P_t$
- Calculate motion feature vector $M_t$
- $F_t = [P_t, M_t]$
- Normalizing $F_t$ with the scaling parameter s to obtain $SF_t$
- Feature selection: $\widehat{SF}_t$
- Classify $\widehat{SF}_t$ using the RBF-SVM
- Update $\theta_t$
if $\theta_t \geq \theta_{min}$ then
    - detected = true
else
    - Shift temporal window by S
    - Retrieve frame t + 1
end if
```

3.5. Action recognition

Real-time activity recognition is achieved using the procedure summarized by algorithm 2. Once the RBF-SVM classifier is trained, the algorithm processes a stream of depth images in order to detect whether a suicide by hanging attempt is underway. First, the positions of the body joints are estimated using the method published in (Shotton et al., 2013). These joints are employed to derive the two sets of features: pose and motion features combined in a single vector F_t. F_t is then scaled and feature selection is applied to generate the observations $\widehat{SF}_t$. These observations are sequentially classified as positive or negative using the RBF-SVM classifier. Suicide detection is based on observing the person's behavior during a sliding temporal window of width Δ_0, which is regularly shifted by a temporal step S at each iteration. Finally, a suicide by hanging attempt is detected if the percentage of positive observations θ_t exceeds a threshold θ_{min}.

4. Experiments

4.1. Dataset

Since there is no public video dataset that can be used to evaluate the proposed system. We thus created our own sequence collection, where 21 persons participated to perform numerous simulations in a room whose dimensions are close to those of prisons. Our system is composed of an RGB-D sensor (Kinect v2) installed in an upper corner, at a distance of approximately 0.30 m from the ceiling, with a tilt angle of 35°. To guarantee an efficient body pose estimation, we suppose that the distance between the candidate and the camera is between 0.5 m and 4.5 m.

To create the dataset, the participants were asked to perform two scenarios:

- In the first scenario, the candidate was invited to simulate a suicide attempt through three main steps: 1) he/she tries to create a hangman's knot using a bed sheet (each person can create the knot in a different way), 2) he/she places the knot in a fixed point in the room, 3) he/she places the knot around the neck.

- During the second scenario, the participant performed some unsuspected actions from different viewpoints with respect to the camera, such as walking in the room, sitting, removing a piece of clothing, wearing clothes, etc.

To learn suicidal behavior, we focus on the action of putting the knot from overhead down to the neck. For each video sequence corresponding to suicide by hanging, we annotated the frames containing this behavior as positive. The frames in the non-suicidal video sequences were annotated as negative. All participants performed suicidal actions, while only the first 15 ones performed non-suicidal actions. The descriptions of the experimental dataset are given in Table1 and Table 2. Our dataset includes a total number of 42 videos. 27 video sequences containing suicide by hanging attempts and 15 video sequences contain only normal activities. In term of frames, our dataset contains 12991 positives frames and 31474 (21393+10081) negative frames.

Table 1: Number of positive frames (suicidal) in each suicidal video sequences. The rest of frames are annotated as negatives.

Sequence index	**# of positive frames / sequence size**	**Sequence index**	**# of positive frames / sequence size**
1	219 / 918	15	302 / 905
2	320 / 895	16	1072 / 1637
3	462 / 1186	17	662 / 1160
4	644 / 909	18	658 / 778
5	464 / 882	19	286 / 758
6	392 / 711	20	520 / 811
7	890 / 1126	21	366 / 753
8	552 / 1338	22	686 / 1483
9	594 / 1054	23	480 / 570
10	250 / 834	24	268 / 395
11	574 / 1989	25	566 / 1079
12	376 / 1086	26	292 / 916
13	298 / 1017	27	630 / 1451
14	168 / 2960	**Total**	**12991 / 29610**

Table 2: Number of negative frames in the video sequences of unsuspected actions.

Sequence index	**# of negative frames**	**Sequence index**	**# of negative frames**
1	1106	9	2084
2	1249	10	1132
3	1370	11	1032
4	2621	12	1339
5	1563	13	971
6	1591	14	1567
7	1666	15	1090
8	1012	**Total**	**21393**

The dataset was collected and research was conducted with approval from the Research Ethics Board (REB) of École de

Technologie Supérieure, Montréal (QC), Canada.

5. Results

In order to test the accuracy and stability of the presented method, we carried out two different tests using the dataset described above.

- In the first test, we use the leave-one-sequence out cross validation procedure to evaluate our system on all sequences in the dataset.
- Since the dataset was unbalanced because the action of interest takes a short time (a few seconds) compared to non suspecious activities, we perform the second test by splitting the dataset into testing data and training data, and rebalance the latter using the SMOTE: Synthetic Minority Over-sampling Technique (Chawla et al., 2002)

For each experiment, we applied the algorithm, depicted in Fig. 3, on the training set in order to extract feature vectors and generate the trained binary classifier. Different simple classifiers have been tried including Linear Discriminant Analysis (LDA), Linear Support Vector Machines (L-SVM), Naive Bayes (NB) and Support Vector Machines with radial basis kernel (RBF-SVM). Note that the optimal results in all experiments have been obtained using only 100 selected features by using the entropy score as a max-relevance criterion in the mRMR algorithm described in the section 3.4. Once the trained binary classifier was obtained, we evaluated it using two strategies. The first one is called *Sequence-based suicide detection* which is based on evaluating the classifier on detecting whether a suicide attempt exists in a video sequence. For this purpose, we used the algorithm 2 where the hyper-parameters were empirically fixed. The sliding temporal window Δ_0 was fixed empirically to 4 seconds and shifted at each iteration by $S = 0.07$ seconds. A suicide action was detected if the threshold θ_t exceeds $\theta_{min} = .75$. The second strategy is *frames-based suicide detection* and it was introduced for the following reason: According to the first strategy, the system can recognize a suicide action in a sequence by correctly classifying only the positive frames in the last temporal window. In other words, the system only correctly classify a few positive frames from the whole positive frames in the suicide sequence. For this, we introduce this test to evaluate the effectiveness of our system in classifying individual frames.

5.1. Leave-one-sequence out

In this test, a leave-one-sequence-out cross validation procedure has been applied. In this way, the system is trained with all sequences except one, which is the one that evaluates the accuracy score. By iterating over all the 42 sequences (27 suicidal sequences and 15 normal sequences), the average success rate is used as a final result. Table 3 shows the results of sequence-based suicide detection using the leave-one-sequence out technique. We can see that the best performance was achieved using the L-SVM

Table 3: Sequence-based suicide detection results using leave-one-sequence out test. We present the results obtained by 4 classifier: Linear Discriminant Analysis (LDA), Linear Support Vector Machines (L-SVM), Support Vector Machines with radial basis kernel (RBF-SVM: σ = 4, C = .5) , Naive Bayes (NB)

	LDA	L-SVM	NB	RBF-SVM
Accuracy (%)	88	**90**	85	83
False alarm (%)	6	**0**	26	0

classifier with 90% of accuracy and 0% of false alarm. Although our system did not make any false alarm, it failed to detect some suicidal attempts. This is can refer to two main reasons: 1) The training dataset is unbalanced such that the number of negative frames is much greater than the number of the positive frames. This allows the classifier to more learn and generalize on the majority class (the negative class) and consequently do not make much false alarm, 2) taking a closer look to the misclassified sequences, it can be seen that some suicidal actions are very similar to the normal activities such as wearing clothes action which usually requires moving the hands from bottom to top around the neck.

Table 4 summarizes the results of classifying individual frames using the leave-one-sequence out. As in the first test, we remark that the best result was obtained using the L-SVM classifier. The small false alarm rate confirms the result obtained in the first test. However, even if the accuracy of classifying individual frames is acceptable with respect to the first test, it can be improved by rebalancing the dataset to allow the classifier an equitable learning on both classes.

Table 4: Frame-based suicide detection results using leave-one-sequence out test.We present the results obtained by 4 classifier: Linear Discriminant Analysis, Linear Support Vector Machines, Support Vector Machines with radial basis kernel (σ = .5, C = .5) , Naive Bayes

	LDA	L-SVM	NB	RBF-SVM
Accuracy (%)	84	**85**	84	**83**
False alarm (%)	10	**6**	13	**4**

5.2. Cross-validation on rebalanced dataset

In this test, We randomly split the available video sequences into a training and a testing sets: 32 video sequences are used to train the proposed system, while the rest (10 sequences) are employed exclusively for testing. As described in Table 1 and Table 2, the number of negative frames are much greater than the number of positive frames. Thus, the classifiers tend to draw a good accuracy on the majority class, having simultaneously a very poor accuracy on the minority class. This is due to the fact that most classifiers pursue minimizing the error rate without taking into account the large difference between the number of cases belonging to each class. To overcome this problem, we propose to use the SMOTE: Synthetic Minority Over-sampling Technique (Chawla et al., 2002) in order to rebalance the training data set by oversampling the positive class (suicidal behavior).

Table 5: Sequence-based suicide detection results using a balanced dataset. We present the results obtained by 4 classifier: Linear Discriminant Analysis, Linear Support Vector Machines, Support Vector Machines with radial basis kernel (σ = 4, C = .5) , Naive Bayes

	LDA	L-SVM	NB	RBF-SVM
Accuracy (%)	90	90	**100**	**100**
False alarm (%)	80	40	**0**	**0**

Table 5 presents the sequence-based suicide detection results obtained using the balanced training data. The best accuracy was obtained using NB (Naive Bayes) and RBF-SVM (SVM with radial basis kernel). We obtained a 100% accuracy with 0% false alarm using 5 normal video sequences and 5 sequences containing suicidal behavior. In Table 6, the frame-based suicide detection results obtained using the balanced data set confirm the effectiveness of our data balancing strategy to improve the system performance.

Table 6: Frame-based suicide detection results using a balanced dataset. We present the results obtained by 4 classifier: Linear Discriminant Analysis, Linear Support Vector Machines, Support Vector Machines with radial basis kernel (σ = 4, C = .5) , Naive Bayes

	LDA	L-SVM	NB	RBF-SVM
Accuracy (%)	89	89	89	**90**
False alarm (%)	9	10	10	**8**

6. Limitations discussion

The Kinect camera provides an informative 3D-skeleton in real-time, which is extremely valuable for activity recognition applications. However, some limitations of such sensor can be figured out and discussed.

From one hand, although that Kinect camera does not require a perfect illumination condition, the direct sunlight can still saturate the sensor, in which case the depth values could not be accurately estimated. Fortunately, prison cells are often not directly "naturally" enlightened owing to the absence of windows. In the case of a naturally illuminated scene, another RGB-based method could be used in parallel(e.g. detecting the presence of a hangmans knot using object detection). From the other hand, even that prison cells are not generally cluttered scenes, the occlusion problem still exists and can affect the reliability of the 3D skeleton detection. In this case, two approaches could be explored: 1) a traditional multi-cameras system based on two Kinect sensors operating in parallel can be used to overcome this problem. We also believe that the use of two cameras, with the appropriate positioning, can resolve the problem of the operation range, which is limited to 1 to 4 meters. 2) A method based on tracking the head and check the presence of a hangman's knot object around the neck can be developed. In fact, the head region is often visible to the camera and its tracking can be achieved by mapping a 2D skeletons on RGB images.

7. Conclusion

We introduced an intelligent surveillance system for detecting suicide by hanging attempts in prisons. Our algorithm exploits the 3D joint's positions acquired in real time by a Kinect camera. We used pose and motion features to represent human movements and model the action of interest. A feature selection technique is applied to speed up our algorithm and improve its generalization capacity. Once the system is trained, we apply an online detection procedure, specifically designed to meet the real-time requirements. Several experiments have been carried out using numerous classifiers in order to evaluate our system. These tests allowed to achieve a high accuracy, especially after rebalancing the training set.

In the future work, we plan to collect more data to really rebalance (without using artificial data) the dataset and increase the testing data. We will also focus on improving the capacity of our system on classifying individual frames which consequently can help the system to detect the suicide attempts in a shorter time. In this work, an important line of work is to model and recognize the action of creating the knot using the strangling object. This can be achieved by combining depth information with the corresponding near infrared and RGB images offered by the Kinect camera, which serves for detecting the knot.

References

Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: Speeded up robust features. Computer vision–ECCV 2006 , 404–417.

Bouachir, W., Noumeir, R., 2016. Automated video surveillance for preventing suicide attempts. 7th IET International Conference on Imaging for Crime Detection and Prevention (ICDP 2016) .

of the Correctional Investigator of Canada, O., 2014. A Three Year Review of Federal Inmate Suicides (2011–2014). Technical Report.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357.

Cheema, S., Eweiwi, A., Thurau, C., Bauckhage, C., 2011. Action recognition by learning discriminative key poses, in: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 1302–1309.

Cook, F.E., 2011. Door suicide alarm. US Patent RE42,991.

Fathi, A., Mori, G., 2008. Action recognition by learning mid-level motion features, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008., pp. 1–8.

Gouiaa, R., Meunier, J., 2015. Human posture recognition by combining silhouette and infrared cast shadows, in: 2015 International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 49–54.

Hayes, L.M., 2013. Suicide prevention in correctional facilities: Reflections and next steps. International journal of law and psychiatry 36, 188–194.

Johansson, G., 1973. Visual perception of biological motion and a model for its analysis. Attention, Perception, & Psychophysics 14, 201–211.

Lee, S., Kim, H., Lee, S., Kim, Y., Lee, D., Ju, J., Myung, H., 2014. Detection of a suicide by hanging based on a 3-d image analysis. IEEE Sensors Journal 14, 2934–2935.

Noonan, M., Ginder, S., 2013. Mortality in local jails and state prisons, 2000-2011, statistical tables. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics Washington, DC.

Parisi, G.I., Weber, C., Wermter, S., 2015. Self-organizing neural integration of pose-motion features for human action recognition. Frontiers in Neurorobotics 9, 3. URL: `https://www.frontiersin.org/article/10.3389/fnbot.2015.00003`, doi:`10.3389/fnbot.2015.00003`.

Parisi, G.I., Wermter, S., 2013. Hierarchical som-based detection of novel behavior for 3d human tracking, in: The 2013 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. doi:`10.1109/IJCNN.2013.6706727`.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence 27, 1226–1238.

Sarafianos, N., Boteanu, B., Ionescu, B., Kakadiaris, I.A., 2016. 3d human pose estimation: A review of the literature and analysis of covariates. Computer Vision and Image Understanding 152, 1 – 20.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R., 2013. Real-time human pose recognition in parts from single depth images. Communications of the ACM 56, 116–124.

Weinland, D., Ronfard, R., Boyer, E., 2011. A survey of vision-based methods for action representation, segmentation and recognition. Computer vision and image understanding 115, 224–241.

World Health Organization, 2007. PREVENTING SUICIDE IN JAILS AND PRISONS. Technical Report.

World Health Organization, year=2014, p., . Preventing suicide: a global imperative.

Yang, X., Tian, Y., 2014. Action recognition using super sparse coding vector with spatio-temporal awareness, in: European Conference on Computer Vision, Springer, Cham. pp. 727–741.

8. Acknowledgment

This work was supported by research grants from the Natural Sciences and Engineering Research Council of Canada, MITACS, and an industrial funding from Aerosystems International Inc. The authors would also like to thank their collaborators from Aerosystems International Inc.