Computer Vision and Image Understanding 197–198 (2020) 102991

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

An attention recurrent model for human cooperation detection

David Freire-Obregón [a,*], Modesto Castrillón-Santana [a], Paola Barra [b], Carmen Bisogni [b], Michele Nappi [b]

[a] Universidad de Las Palmas de Gran Canaria, Spain
[b] Università degli Studi di Salerno, Italy

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

ABSTRACT

User cooperative behaviour is mandatory and valuable to warranty data acquisition quality in forensic biometrics. In the present paper, we consider human cooperative behaviour in front of wearable security cameras. Moreover, we propose a human cooperation detection pipeline based on deep learning. Recently, recurrent neural networks (RNN) have shown remarkable performance on several tasks such as image captioning, video analysis, or natural language processing. Our proposal describes an RNN architecture with the aim at detecting whether a human is exhibiting an adversarial behaviour by trying to avoid the camera. This data is obtained by analysing the noise patterns of human movement. More specifically, we are not only providing an extensive analysis on the proposed pipeline considering different configurations and a wide variety of RNN types, but also an ensemble of the generated models to outperform each single model. The experiment has been carried out using videos captured from a mobile device camera (GOTCHA Dataset) and the obtained results have demonstrated the robustness of the proposed method.

1. Introduction

During the last two decades, public places have experienced a rapid increase of deployed surveillance cameras. The impact of terrorist acts against our society does not only affect our sensibility, but also our economy, our political ideas, and even jeopardize our lifestyle. Regarding to this issue, camera surveillance can be considered as a tool of situational crime prevention (Stutzer and Zehnder, 2013). As a result, humans have learned to live under constant observation in public areas. These devices have been installed in airports, train stations, commercial facilities, hospitals, schools, etc. No one can deny that safety is a trending topic in our society and everyone is likely to be recorded on video in the aforementioned public spaces.

Most conventional surveillance systems rely on security personnel to monitor and detect suspicious activities using surveillance equipment. The increase in the number of these cameras carry a new issue, the overload of visual data. In the majority of the cases, the security personnel is the most accurate way of filtering this data, which results in high costs and low efficiency.

In this sense, video surveillance has become one of the most relevant safety methods to prevent undesirable situations. From a technical point of view, the main goal of the research on the surveillance video techniques is to effectively extract and analyse information from a large amount of unstructured data. This analysis must be done automatically, tracking and identifying suspicious behaviours in order to ask for human supervision.

A wide variety of reliable methods for automatically analysing surveillance videos have been recently developed. Traffic controlling (Vinayaga-Sureshkanth et al., 2018; Sochor et al., 2019), crime prevention (Khan et al., 2018; Piza, 2018), and security monitoring (Chaaraoui et al., 2013; Wang, 2013; Singh et al., 2018) provide some of the most interesting and practical applications. Furthermore, much attention has been drawn in the computer vision community due to their high impact on the security market (Yu and Moon, 2009; Oluwatoyin and Wang, 2012; Joshi et al., 2016; Aitfares et al., 2016; Alshalawi and Alghamdi, 2017).

However, surveillance cameras are not static devices anymore. The evolution of technology through the miniaturization of capture devices and storage technologies has created new opportunities for the law enforcement agencies. Many police departments have successfully administered the use of body-worn cameras from this advancement of technology. There is significant potential for wearable technology to enhance effectiveness and safety by gathering additional data and providing critical information to officers (Brown and Fan, 2016). As a result, real-time analysis is needed to assist these agents in several security issues such as suspicious object detection and anomaly detection in given videos.

Regarding anomaly detection, psychological studies analysed the effect of camera surveillance in correlation of certain of deviant behaviour (Hoffmann, 2011; Bateson et al., 2006). The results show

* Corresponding author.
E-mail address: david.freire@ulpgc.es (D. Freire-Obregón).

https://doi.org/10.1016/j.cviu.2020.102991
Received 21 October 2019; Received in revised form 7 May 2020; Accepted 25 May 2020
Available online 2 June 2020

that the presence of surveillance cameras lead people to be more cooperative and exhibit prosocial behaviour. These results support the hypothesis that the way the camera is presented indeed influence people's behaviours, i.e., propensity for deviant behaviour decreases as the risk of being caught increases. Moreover, our main contribution is the development of an effective pipeline to detect adversarial human behaviour when facing wearable cameras. This adversarial behaviour is defined in terms of showing a non-cooperative attitude towards the wearable cameras, i.e. trying to fool the system (Micheloni et al., 2009).

As a consequence, the major contributions of this study are as follows: (1) to tackle human adversarial behaviour problem, (2) the design of an efficient attentive recurrent architecture, (3) the evaluation of different recurrent architectures configurations, and finally, (4) the successful application of the proposed approach to boost the cooperative/non-cooperative detection performance.

The paper is organized into six sections. The next section discusses some related work of the state of the art. In Section 3, data preprocessing and the attentive recurrent architecture are described. The experimental setup in addition to the classification experiments are reported in . Then, an ensemble of several configurations of the proposed pipeline is addressed in Section 5. Finally, conclusions are drawn in Section 6.

2. State of the art

In the previous section, surveillance cameras were defined as capture devices that provide useful footage for forensics purposes. The information provided by this technology is very helpful as evidence for legal issues (Mahmood Rajpoot and Jensen, 2015). In this section, we address the way this footage has been used to tackle security issues in the past.

Motion detection is an issue usually tackled by almost every visual surveillance system. The task is to segment regions corresponding to moving objects from the rest of the image. The subsequent processes depends on both the nature of the segmented object and the addressed problem. Moreover, the segmented regions in image sequences can correspond to moving objects such as humans or vehicles. Consequently, these detected moving regions provide a focus of attention for later processes such as tracking and behaviour analysis as only these regions need to be considered and further investigated. In addition, there is a notable intersection between tracking algorithms and motion detection during processing. Indeed, tracking over time typically involves matching moving objects in consecutive frames using features like points or lines. Security approaches can be classified into two major groups depending on the purpose of the conducted research.

The first group gathers all the methods that require to identify abandoned objects. Furthermore, abandoned detection is one of the most important tasks in automated video surveillance systems. In this regard, approaches based on background subtraction have shown a remarkable robustness in complex real-world scenarios (Wen et al., 2009; Li et al., 2010). Background maintenance and static foreground object detection are the main challenges in such approaches. For the background maintenance, a Mixture of Gaussians is usually considered to model both background and foreground for each individual pixel (Li et al., 2010; Tian et al., 2011). On the other hand, the algorithms for identifying a static foreground involves constructing double-background models for detecting a static foreground (see Porikli et al., 2008; Evangelio et al., 2011).

The second group gathers all the methods that involve behaviour understanding. In some circumstances, it is necessary to analyse the behaviours of people and determine whether their behaviours are normal or abnormal (Ko, 2011). Furthermore, the purpose of these methods is to produce a high-level description of interactions and behaviours between humans. In this paper, we are going to categorized these descriptions as indirect and direct descriptions. Indirect descriptions of human behaviour can be obtained by the observation of external entities guided by humans such as vehicles (Song et al., 2018), videogames characters (Lamb et al., 2018), etc. Conversely, direct descriptions can be obtained by observing humans behaviour directly.

Recently, Alexandrie (2017) found out that video surveillance can reduce crime, showing reductions ranging from 24%–28% in public streets and urban subway stations where surveillance cameras were placed. Unfortunately, traditional static surveillance cameras contain blind spots such as alleyways, where some offenders can avoid being recorded. However, wearable cameras are dynamic and offenders need to perform deliberate actions to not get captured by these devices. This paper address this issue, suggesting to obtain a direct description (cooperative/non cooperative) by observing humans behaviour under different lighting conditions. In the literature, cooperative can be defined in multiple ways, such as the ability to provide help to accomplish a task such in robotics (Abu Bakar et al., 2009) or the ability to provide some human-hints to assist a computer vision algorithm (Nagai et al., 2017). However, we define this concept in terms of security as the detection of normal (cooperative) or abnormal (non cooperative) behaviour towards a wearable cam.

Lately, traditional Deep Learning (DL) studies about human activity have been focused on human action recognition. Ji et al. (2013) developed a 3D convolutional neural network (CNN) model for action recognition in surveillance videos. More recently, Ullah et al. (2018) proposed a framework for activity recognition in surveillance videos captured over industrial systems. They combined a CNN based human saliency features, a FlowNet2 CNN model, and a multi-layer long short-term memory (LSTM) to learn long-term sequences in the temporal optical flow features for activity recognition. Sharma et al. (2015) proposed multi-layered RNNs with LSTM units which are both spatially and temporally deep. Also Yeung et al. (2015) addressed the human action recognition by defining a novel variant of LSTM deep networks for modelling these temporal relations via multiple input and output connections. Another interesting research was conducted by Yu and Qin (2018), reaching an accuracy of 93.79%. They proposed a bidirectional LSTM (BiLSTM) structure for human activity recognition using time series, collected from a smartphone.

Precisely, the present work focuses on designing an attentive recurrent neural network in order to have an effective tool for cooperative/non cooperative behaviour classification. The proposal is based on an attentive recurrent architecture to provide good discrimination among different human behaviour towards a wearable device.

3. Proposed pipeline

As can be seen in Fig. 1, a sequential pipeline divided into two main blocks is proposed and evaluated. The first block preprocesses the raw data in order to feed the second block (the RNN model). This figure also shows how the data preprocessing block can be split into four sublocks (labelled as A, B, C and D). Firstly, once the image is captured by the acquisition device (Fig. 1-A), the algorithm proposed by Cao et al. (2018) is applied in order to obtain a set of keypoints all around the body based off the COCO pose output (Fig. 1-B). Legs keypoints are not taken into consideration to provide a more flexible model (see Fig. 2). Secondly, different configurations of distances between these keypoints are analysed to find the most relevant in order to solve the proposed problem (Fig. 1-C). Thirdly, a bucketing technique (Fig. 1-D) is applied to reduce the amount of information processed by the RNN. Finally, this RNN is applied to train, validate, and test the preprocessed data.

3.1. Skeleton representation

In the last few years, the skeleton representation has been used as a recurrent data source to solve the human action recognition issue (see Pang et al., 2013; Simonyan and Zisserman, 2014a; Song et al., 2017; Wang et al., 2017). As already mentioned above, we have considered the human body model provided from the Cao et al. (2018) algorithm. For robustness purposes, we only use a subset of keypoints

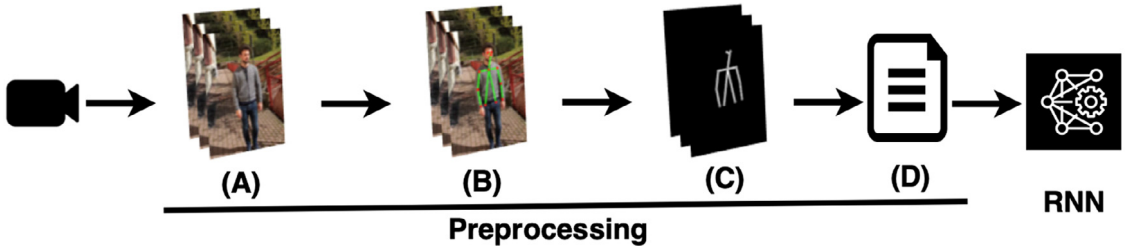

Fig. 1. The proposed pipeline for the cooperative/non cooperative problem comprises two main blocks: the preprocessing block and the attentive RNN application. The preprocessing block implies: the image acquisition (A), the keypoints extraction (B), the distance skeleton configuration selection, (C) and the application of the bucketing algorithm (D).

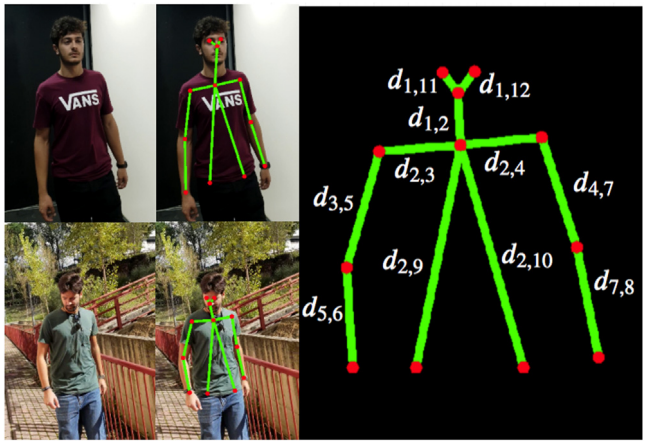

Fig. 2. Distance skeleton extraction. In the left side of the figure we can appreciate two different subjects and how the provided keypoints (red circles), known as joints, are used to generate the distance skeleton. Moreover, different skeleton configurations were evaluated in our experiments. The upper left image shows a cooperative behaviour while the bottom left image shows quite the opposite, a non-cooperative behaviour.

from the COCO pose output. Therefore, these keypoints can be defined as a set of $n = 12$ joints in 3-D (see the red circles in Fig. 2). Each joint can be denoted as $v_i = [x_i, y_i, z_i]^T$, where (x_i, y_i) are the 2-D spatial coordinates and z_i is the confidence score in the range $[0, 1]$. Thus, the articulated pose configuration of the human body can be represented by concatenating the three coordinates of all joints v_i:

$$V = [v_1, v_2, v_3, \dots, v_n]^T \tag{1}$$

The proposed approach does not consider the joints by itself but the variation of the distance between joints at every moment. This distance can be formulated as follows:

$$d_{p,q}(x, y) = \| v_p(x, y) - v_q(x, y) \| \tag{2}$$

where $d_{p,q}(x, y)$ is the spatial distance between joints $v_p(x, y)$ and $v_q(x, y)$ respectively. The resulting confidence can be formulated as follows:

$$d_{p,q}(z) = (v_p(z) * v_q(z))/((v_p(z) + v_q(z))/2) \tag{3}$$

where $d_{p,q}(z)$ is the new probability considering both, $v_p(z)$ and $v_q(z)$. Therefore, the joints distance configuration of the human body can be represented as follows:

$$D = [d_{1,2}, d_{1,11}, d_{1,12}, d_{2,3}, \dots, d_{m,n}]^T \tag{4}$$

Moreover, as can be appreciated in Fig. 2, the joints $V(x, y)$ are represented by red circles, while the distances $D(x, y)$ are represented by green lines. In this sense, the D configurations considered in our experiments rely on connected joints and are defined as follows:

- Neck configuration: $D_{neck} = [d_{1,2}]$
- Vertical configuration: $D_{ver} = [d_{1,2}, d_{2,9}, d_{2,10}]$
- Horizontal configuration: $D_{hor} = [d_{1,2}, d_{2,3}, d_{2,4}]$
- Full configuration: $D_{full} = [D_{ver} \cup D_{hor}]$

The main idea is not only detect obvious camera avoidance (i.e., using the hands to hide the face), but also subtle camera avoidance. Being able to train a model that detects this second kind of avoidance without consideration of the arms, provides more robustness for a wider set of situations. Eq. (4) can be extended to each video frame and the number of frames may vary between videos. The training of RNN for large videos is computationally expensive and the sequential nature of the recurrent process forbids the parallelization of the training over input dataframes. A robust optimization requirements to work on large batches of data and training time as well as classification performance can vary strongly depending on the choice of how batches were put together.

3.2. Bucketing algorithm

The bucketing algorithm provides a framework where each bucket represents a range of data D from different dataframes. Let $S = \{D_1, D_2, D_3, \dots, D_n\}$ be the set of distances for a video sequence of n frames. Bucketing is defined as the process when all sequences are clustered into r buckets of a fixed length l, where r is some small positive integer number (see Fig. 4). In a more general mathematical sense, the bucketing algorithm provides the following output for each video sequence (see Fig. 5):

$$X = [x_1, x_2, x_3, \dots, x_r]^T \tag{5}$$

Accordingly, x_i denotes the i-bucket computed as:

$$x_i = \Lambda(D_{[j,k]}) \qquad \forall i, j \in n \wedge i \neq j \tag{6}$$

where $\| k - j \|$ is the bucket length (l). The $\Lambda(D_{[j,k]})$ function provides a tuple $\langle \overline{m}, \sigma, \overline{m} \pm \sigma, P_i \rangle / i = [0, 25, 50, 75, 100]$ of the raw data but also of the differential data obtained from the differences between the successive frames. Finally, raw and differential data are normalized in order to feed the proposed attentive recurrent architecture. Bucketing techniques are usually considered for sequential problems, such as natural language processing (NLP), to increase the training speed for tasks where a length of the input sequence may vary significantly (Khomenko et al., 2016). We have used 4 GeForce GTX 1080ti to train our models at its full capacity. Increasing the number of buckets affects the time in order of magnitude 1% per bucket. For example, a model using 50 buckets as input was trained twice as fast as the same model using 150 buckets as input.

3.3. Attentive recurrent architecture

RNNs are known as a class of neural network models that have a self-connected hidden layer. It is able to process a sequence of arbitrary length by recursively applying a transition function to its internal hidden state vector h_t of the input sequence. Moreover, the activation of the hidden state h_t at time-step t is computed as a function f of the current input x_t and the previous hidden state h_{t-1}:

$$h_t = \begin{cases} 0, & \text{if } t = 0. \\ f(h_{t-1}, x_t), & \text{otherwise.} \end{cases} \tag{7}$$

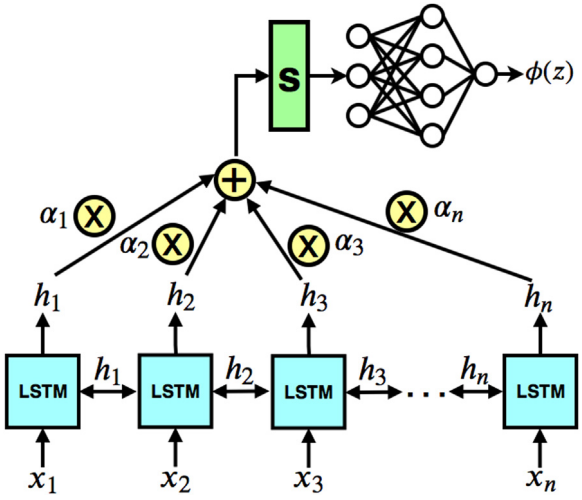

Fig. 3. Schematic of the proposed architecture with a feed-forward attention mechanism. Vectors in the hidden state sequence h_t are fed into the differentiable function $Q(h_t)$ to produce a probability vector α. The vector S is computed as a weighted average of h_t, with weighting given by α. The resulting vector S feeds a fully connected layer to generate the classification output $\phi(z)$.

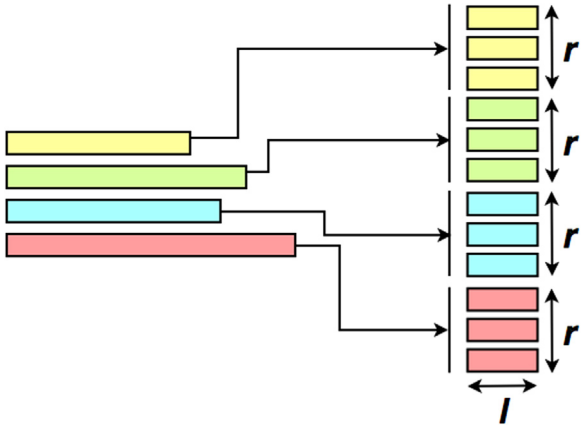

Fig. 4. Bucketing in a sequential model. The number of dataframes per video may vary. The bucketing algorithm provides the possibility to fix the amount of information per video to feed the Attentive RNN.

Consequently, the recurrent connection creates an internal state of the network, allowing it to record internal states to consider a past context. An important issue of the standard RNNs refers to a limited contextual information range. It is hard to learn long-term contextual dependencies. The recurrent connection causes the input influence to either decay or blow up exponentially, commonly known as the vanishing gradient problem (Graves et al., 2009). The LSTM network was proposed by Hochreiter and Schmidhuber (1997) to specifically address the issue of learning long-term dependencies. The LSTM maintains a separate memory cell inside it that updates and exposes its content only when needed. Moreover, LSTM carries data from various steps and each cell step is capable of adding and removing information from this data while processing a new sequential input. Therefore, LSTM cells have layers called "gates" which will allow information to be "forgotten" or "perpetuated" to next steps/cells. Goodfellow et al. (2016) have identified the following gates:

- The forget gate as $f_t = \phi(W_f \cdot [h_{t-1}, x_t] + b_f)$, where W_f denotes the weights of the forget gate, b_f denotes the bias and ϕ denotes the logistic sigmoid function. This gate forgets information that is no longer necessary.
- The input gate as $i_t = \phi(W_i \cdot [h_{t-1}, x_t] + b_i)$. This gate allows to introduce new information considering the current step x_t and the previous output h_{t-1}.
- The output gate as $o_t = \phi(W_o \cdot [h_{t-1}, x_t] + b_o)$. This gate controls the output of the cell and the internal memory state.

Intuitively, considering the previous gates, the new tth values are computed as follows:

$$C_t = f_t * C_{t-1} + i_t * \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{8}$$

where the internal hidden state vector h_t is computed as:

$$h_t = o_t * \tanh(C_t) \tag{9}$$

In this paper, we use BiLSTM (Schuster and Paliwal, 1997). This type of RNN combines an LSTM that moves forward and backwards through time.

Generally, not all the human movements contribute equally to the representation of our addressed problem. We have introduced a movement attention mechanism to capture the distinguished influence of the movements (described as X in the previous subsection) on the cooperative/non cooperative issue, and then form a dense vector considering the weights of different movements vectors (Bahdanau et al., 2014). From a mathematical perspective, attention can be defined through a context variable S:

$$S = \sum_{t=1}^{T} \alpha_t * h_t \tag{10}$$

where T denotes the total number of time steps in the input sequence and α_t is a weight computed at each time step t for each state h_t. As demonstrated in Fig. 3, we are considering a straightforward simplification to the attention mechanism which would allow it to be used to produce a single vector S from an entire sequence (Raffel and Ellis, 2015). In Eq. (10), attention generates a fixed-length embedding S of the input sequence by computing an adaptive weighted average of the state sequence h_t. The value of the parameter α can be formulated as follows:

$$\alpha_t = \frac{exp(Q(h_t))}{\sum_{k=1}^{T} exp(Q(h_k))} \tag{11}$$

where the learnable function $Q(h_t)$ can be defined as:

$$Q(h_t) = \tanh(W_Q \cdot h_t + b_Q) \tag{12}$$

Usually, the meaning of the output of a RNN at each time step depends on the addressed problem. We can appreciate in Fig. 3 that our proposal can be seen as a sequence to one approach. The input represents the data sequences addressed in the previous subsection while the output is one single label (cooperative/non cooperative) computed by the BiLSTMs, the attention mechanism, and a densely connected classifier as previously described.

4. Results

This section is divided into two subsections related to experimental issues: setup and results. The first subsection describes not only the considered dataset for our particular problem but also different technical details such as the considered metric. The achieved results are summarized in the second subsection.

4.1. Experimental setup

In order to solve the proposed issue, we have used the GOTCHA Dataset for our experiments. Barra et al. (2019) introduced this unique dataset of videos taken from a mobile device, for the purpose of development of mobile-devices-based forensic and biometric methods. Contrary to other popular datasets such as SBU (Yun et al., 2012) or

Fig. 5. Extracted points in a GOTCHA Dataset sample video. The reader may observe that the camera is not static, it moves with the subject while trying to capture a clear image of the individual. As a consequence, keypoints are not available for every frame. We have developed three steps to deal with this situation: (a) only distances are considered, (b) missing values are handled by applying interpolation, and (c) the sequence of keypoints are normalized considering $d_{1,2}$.

Fig. 6. The GOTCHA Dataset provides several challenging situations. The top frames of the figure show different perspective from the acquisition device of a same clip. The bottom frames shows not only different scenarios but also several situations such as subject shift (bottom left image) or blurry images (bottom centre image).

NTU (Shahroudy et al., 2016; Liu et al., 2019), the GOTCHA Dataset consists of 6 clips per subject covering cooperative and non-cooperative mode including indoor with artificial light, indoor without lights (flash camera on), and outdoor with sun light. The total number of recorded subjects is 62. The dataset contains over 372 short clips (around 10 seconds each) where subjects are captured while they walk exhibiting different behaviours in front of the camera: avoiding the camera (not cooperative) or just ignoring it (cooperative). A remarkable feature about this dataset is that the camera is not static, it moves with the subject while trying to capture a clear image of the individual. The dataset and the preprocessed data are available for research purposes.[1] As it was previously mentioned, the dataset provides not only videos of indoor and outdoor scenes acquired under widely comparable conditions, but also videos taken under different lighting conditions, i.e., all dark only with the camera flash light on. The acquisition device was the rear camera of a Galaxy Samsung S9+ smartphone. The bottom of Fig. 6 shows images of three subjects for the three previously commented environments.

For each experiment, train and test data are chosen randomly and the results are averaged after considering 5-fold cross validation. However, we split the dataset by subjects to address the cooperative/not cooperative problem. Our purpose by doing this split was to avoid creating bias to our model with the intrinsic gesture recognition of a same person between videos.

In terms of metric, the Matthews Correlation Coefficient (MCC) has been chosen to train the models described in Section 4.2 (Matthews, 1975). Recently, Chicco (2017) claimed that the MCC is more informative than other confusion matrix measures (such as F1-score and accuracy) in evaluating binary classification problems. Chicco argued that the MCC considers the balance ratios of the four confusion matrix categories: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). The MCC can be computed from the confusion matrix as follows:

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TN + FN) * (TN + FP) * (TP + FN)}} \quad (13)$$

The MCC has a range of $[-1, 1]$ where -1 indicates a completely wrong binary classifier and 1 indicates a completely correct binary classifier. As can be seen in Eq. (13), if any of the four sums in the denominator is zero, the denominator can be arbitrarily set to one. The result is a MCC of 0, which can be shown to be the correct limiting value. Moreover, MCC exhibits a more restrictive behaviour than other metrics such as accuracy or recall. For instance, if a binary class problem is balanced and all the samples are classified as only one class, the accuracy will be 50% while the MCC is 0, not 0,5.

4.2. Experimental results

We conducted a set of experiments to validate the effectiveness of the described proposal, henceforth known as Att-RNN. These experiments took place through a grid search considering different parameters. The most relevant parameters were the number of buckets, the joints distance configuration, and the studied architecture.

As we argued in Section 3.2, the number of buckets per video (r) can be fixed providing different perspectives over the data. Using a higher number of buckets where the density of the underlying data points is low can reduce the noise due to sampling randomness. Using a lower number of buckets where the density is high gives greater precision to the density estimation.

The second relevant parameter related to our proposal is the joints distance configuration (D). In this sense, we described different possible configurations in Section 3.1. Four different configurations were considered: neck configuration: (D_{neck}), vertical configuration (D_{ver}), horizontal configuration (D_{hor}) and full configuration (D_{full}).

Finally, a set of different state of the art architectures is needed to compare our results. These approaches were tested under the same circumstances as our proposal. Furthermore, these approaches are as follows:

[1] https://gotchaproject.github.io

Table 1
Best result obtained from each considered approach. The Table is organized in terms of the number of buckets (r).

Number of buckets (r)	Approach	$D_?$	MCC	Accuracy
50	LSTM	D_{hor}	0,920	95,96%
	CNN-LSTM	D_{hor}	0,811	90,05%
	ConvLSTM	D_{hor}	0,903	95,16%
	BiLSTM	D_{hor}	0,915	95,69%
	Att-RNN	D_{hor}	**0,921**	**95,97**%
100	LSTM	D_{hor}	0,909	95,43%
	CNN-LSTM	D_{hor}	0,877	93,82%
	ConvLSTM	D_{hor}	0,919	95,96%
	BiLSTM	D_{hor}	0,909	95,43%
	Att-RNN	D_{hor}	**0,952**	**97,58**%
150	LSTM	D_{hor}	0,893	94,62%
	CNN-LSTM	D_{hor}	0,915	95,69%
	ConvLSTM	D_{hor}	0,892	94,62%
	BiLSTM	D_{hor}	0,903	95,16%
	Att-RNN	D_{hor}	**0,941**	**97,04**%
Raw Pixels	VGG16	–	0,709	76,62%

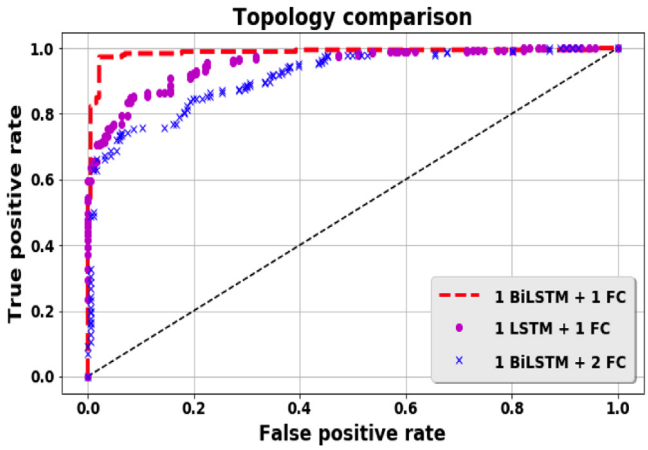

Fig. 7. Topology comparison considering the Att-RNN approach when $r = 100$. Three different topologies are shown in this graph. Considering either 2 fully connected (FC) layers on top of our attention proposal or a unidirectional LSTM as base layer tend to overfit. The optimal accuracy occurs when 1 BiLSTM layer and 1 FC layer are combined in our approach as base and output layers respectively.

- LSTM. This is a classical recurrent neural network architecture widely used in the field of deep learning. Usually this approach is combined with other techniques. For instance, the previously commented CNN-LSTM combines the CNN and the LSTM. As aforementioned, a multi-layer LSTM was combined several models and techniques by Ullah et al. (2018) to learn long-term sequences in the temporal optical flow features for activity recognition.
- CNN-LSTM. Usually, a CNN considers two dimensional data, but it can be also adapted to work with less or more dimensions. Moreover, CNN have shown a remarkable performance dealing with one-dimensional sequence data such as univariate time series data. Recently, Pan et al. (2018) have used this same technique for bearing fault diagnosis and the results were quite impressive. The CNN model is used in a hybrid model with an LSTM backend where the CNN is used to interpret subsequences of input that together are provided as a sequence to an LSTM model to interpret.
- Convolutional LSTM. This approach is somewhat related to the CNN-LSTM. However, in this case, the convolutional reading of the input takes place directly into each LSTM unit (Shi et al., 2015). This model, also known as ConvLSTM, was developed for reading two-dimensional spatial–temporal data, but can be adapted for use with univariate time series data.
- Bidirectional LSTM. This approach allows the LSTM model to learn the input sequence both forward and backwards and concatenate both interpretations. This approach has been previously tested for human activity recognition using time series collected from a smartphone (Yu and Qin, 2018).

Consequently, these approaches have been adapted to tackle the cooperative/non-cooperative problem and added to the proposed pipeline shown in Fig. 1 as the RNN block. Table 1 shows the best results obtained on the test set when we conducted the experiment. The table presents not just the MCC value, but also the accuracy obtained for the best model of each approach. As a result, all approaches have exhibit remarkable performance following the proposing pipeline. However, our approach (Att-RNN) outperforms the others.

Two different issues must be addressed in order to analyse these results. First, the table is divided by the number of considered buckets. A low number of buckets can mean a loss of information. On the other hand, a high number of buckets does not guarantee high accuracy because brief but high data variations can affect the model. There must be an adjustment in order to find the optimal value. In our case, 100 buckets turns out to be the optimal threshold for this parameter. Contrary to the Att-RNN model, the CNN-LSTM approach improves along with the number of buckets (+10% between 50 and 150 buckets).

Second, the best result was provided when the horizontal joints distance configuration (D_{hor}) was considered, with an accuracy of 97,58%. Furthermore, the trend for the rest of experiments was not to consider the most informative configuration (D_{full}), but to consider the horizontal configuration (D_{hor}). This is very relevant to the purpose of this experiment. This means that the subject's shoulders and neck provide more relevant information to tackle our problem over any other joints distance configuration.

We also have developed a raw pixels-based approach concentrating on the raw pixels of the image instead of the skeleton. In this case, VGG16 network is considered to extract the appropriate features to classify the dataset (Simonyan and Zisserman, 2014b). Frames are processed individually by the convolutional base of this pre-trained network. Then, a descriptive analysis is performed in order to obtain statistical measures for each video sequence. Finally, these measures are considered in order to fit a random forest model. This approach shown at the end of Table 1 can be considered as a baseline to compare our results.

As we stated in Section 3.3, we are using a BiLSTM network as the attentive base and a densely connected classifier fed by the context variable S on top of it (see Fig. 3). During the grid search, multiple configurations were tested. We considered classical LSTM layers but also BiLSTM layers as base. Additionally, multiple FC layer configurations on top were considered, changing the number of units per layer and adding regularization (dropout). Fig. 7 shows the ROC curve when different topologies are tested for the Att-RNN model. The effect of applying different layers to the architecture can be appreciated. To this end, we modified not only the type of RNN base layer but also the number of FC layers. Contrary to Song's proposal, using BiLSTM has shown a better performance than considering a LSTM as a base layer (Song et al., 2017). Clearly, a simple FC proposal outperforms more complex configurations to address our problem because this kind of signal is very sensitive and rapidly tends to overfit when multiple layers are considered.

We conducted a last experiment to show the most relevant human parts for the cooperative/non-cooperative detection. As it can be appreciated in Fig. 8, the D_{hor} configuration outperforms any other one, including the raw pixels experiment. Moreover, an alternative configuration not addressed in Section 3.1 is introduced to show that more information does not mean a better classifier in case of considering the skeleton. The signal labelled as "$D_{full} + Legs$" is the attention classifier

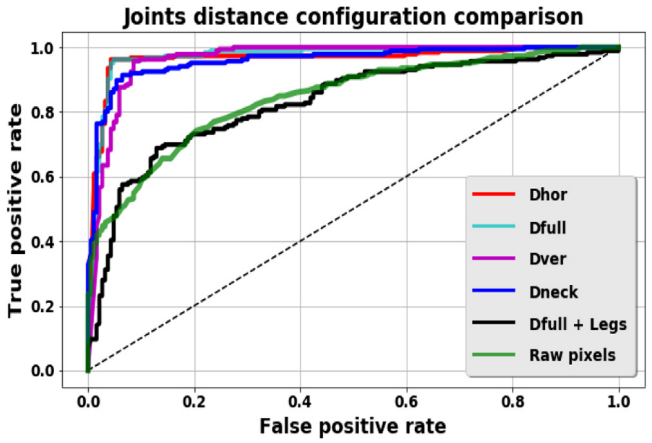

Fig. 8. Joints distance configuration (D) comparison considering the Att-RNN model when $r = 100$. The different configurations described in Section 3.1 are shown in this graph with an extra configuration labelled as "Dfull + Legs" along with the raw pixels experiment previously described. As it can be appreciated, these two experiments have shown a similar performance.

Table 2
Ensemble results for several length combinations.

N_{app}	App_m	r	W_m	MCC	Accuracy
2	Att-RNN	100	0,642	0,957	97,84%
	ConvLSTM	150	0,358		
3	Att-RNN	150	0,562	0,963	98,12%
	Att-RNN	100	0,251		
	BiLSTM	50	0,187		
4	Att-RNN	100	0,371	0,969	98,39%
	CNN-LSTM	150	0,259		
	ConvLSTM	100	0,222		
	ConvLSTM	150	0,148		

response when all the possible information provided by the Cao et al. (2018) algorithm is considered. Although this configuration provides more information than any other configuration, the results are worse than the analysed ones.

5. Diversity-performance trade-off

Finally, we conducted an experiment considering the combination of approaches showed on Table 1. The main motivation was to exploit the dependence between the base learners. Our weighted average ensemble is an approach that allows multiple models to contribute to a prediction in proportion to their trust or estimated performance. The equation is defined as follows:

$$D = \sum_{m=0}^{N_{app}} W_m * App_m \qquad \forall W_m \in [0, 1] \tag{14}$$

where W_m denotes the approach m weight, the App_m denotes the approach m decision ($\mathbb{R} \in [0, 1]$) and N_{app} is the number of considered approaches for the ensemble.

Table 2 shows the best ensemble approaches for several length combinations. The columns of this table represent the approach used for the combination (App_m), the number of buckets (r), the approach weight (W_m) in the decision making process and, the ensemble approach MCC and accuracy. As shown by the table, each ensemble slightly improves our previous results. Our best approach (Att-RNN model when $r = 100$) is included in all the ensemble approaches and the attention proposal is the most highly valued approach due to the weights W_m. A further analysis revealed the existence of diversity-performance trade-off between these models. As Michailidis (2017) stated, the diversity within the combined models is more important in securing a better generalization in the test data than having on average stronger but more correlated models within the ensemble. Moreover, our study fit this theory:

- Firstly, the CNN-LSTM model when $r = 50$ is the weakest classifier by far on Table 1 and is not considered for any combination, only the strongest ones are.
- Secondly, the correlation between the selected stronger models is lower. For instance, the ConvLSTM approach exhibits a sightly lower performance than any other approach when $r = 150$ (see Table 1). However, the correlation between this model and the Att-RNN model (when $r = 100$) is weaker than our proposed approach and any other model when $r = 150$ with a Pearson Correlation Coefficient (PCC) of 0.901. Furthermore, these other models show a PCC that goes from 0.911 to 0.958. Due to this trade-off, the ConvLSTM approach is combined with our approach when $N_{app} = 2$.

6. Conclusions

In this paper, we presented a study based on deep learning to detect human adversarial behaviour using mobile devices. For this reason, we have conducted several experiments considering a complex pipeline that combines different types of RNNs on a dataset of videos acquired from a mobile device. The contribution represents an interesting challenge since there is not much research on this kind of dataset for human behaviour.

As a consequence, our contribution operates from implementing attention on a BiLSTM base to identify whether human behaviour is being natural or trying to avoid a camera. Our findings have shown interesting insights over the behaviour. For instance, when a human tries to avoid a camera, the upper torso part (shoulders and neck) provides the most relevant information. Another interesting insight is provided by the fact that the system operates well enough under different lighting conditions. Indeed, our proposal is sufficiently complex to effectively distinguish between cooperative/non cooperative classes with 97.58% of accuracy. We have also discussed how varying the network topology or how tuning the hyperparameters can affect the model accuracy. In this sense, scalability has been evaluated and increasing the number of layers does not seem to improve a simpler model. Finally, we have shown an ensemble approach combining models that outperforms any of the individual performance of the combined models with a 98.39% of accuracy.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was partially funded by the Universidad de Las Palmas de Gran Canaria grant ULPGC2018-08.

References

Abu Bakar, S.B., Ikeura, R., Salleh, A.F.B., Yano, T., 2009. A study of human-human cooperative characteristic in moving an object. In: ICCAS-SICE. IEEE Computer Society, pp. 1158–1163.

Aitfares, W., Kobbane, A., Kriouile, A., 2016. Suspicious behavior detection of people by monitoring camera. In: 5th International Conference on Multimedia Computing and Systems. ICMCS, pp. 113–117.

Alexandrie, G., 2017. Surveillance cameras and crime: a review of randomized and natural experiments. J. Scand. Stud. Criminol. Crime Prev. 18 (2), 210–222.

Alshalawi, R., Alghamdi, T., 2017. Forensic tool for wireless surveillance camera. In: 2017 19th International Conference on Advanced Communication Technology. ICACT, pp. 536–540.

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. CoRR abs/1409.0473.

Barra, P., Freire-Obregón, D., Bisogni, C., Castrillón-Santana, M., Nappi, M., 2019. Gender classification on 2D human skeleton. In: 3rd IEEE International Conference on Bio-Engineering for Smart Technologies, pp. 123–127.

Bateson, M., Nettle, D., Roberts, G., 2006. Cues of being watched enhance cooperation in a real-world setting. Biol. Lett. 2, 412–416.

Brown, L.M., Fan, Q., 2016. Enhanced face detection using body part detections for wearable cameras.. In: ICPR. IEEE, pp. 715–720.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y., 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. ArXiv Preprint arXiv: 1812.08008.

Chaaraoui, A.A., Climent-Pérez, P., Flórez-Revuelta, F., 2013. Silhouette-based human action recognition using sequences of key poses. Pattern Recognit. Lett. 34 (15), 1799–1807.

Chicco, D., 2017. Ten quick tips for machine learning in computational biology. BioData Min. 10 (1).

Evangelio, R.H., Senst, T., Sikora, T., 2011. Detection of static objects for the task of video surveillance. In: IEEE Workshop on Applications of Computer Vision. IEEE Computer Society, pp. 534–540.

Goodfellow, I.J., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press, Cambridge, MA, USA.

Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., Schmidhuber, J., 2009. A novel connectionist system for unconstrained handwriting recognition. IEEE Trans. Pattern Anal. Mach. Intell. 31 (5), 855–868.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Hoffmann, C., 2011. The Influence of Different Forms of Camera Surveillance and Personality Characteristics on Deviant and Prosocial Behaviour (Bachelor thesis Psychology).

Ji, S., Xu, W., Yang, M., Yu, K., 2013. 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 35 (1), 221–231.

Joshi, T., Aarya, H., Kumar, P., 2016. Suspicious object detection. In: 2nd International Conference on Advances in Computing, Communication, Automation. ICACCA (Fall), pp. 1–6.

Khan, A., Rehman, S., Waleed, M., Khan, A., Khan, U., Kamal, T., Afridi, S.K., Marwat, S.N.K., 2018. Forensic video analysis: Passive tracking system for automated person of interest (POI) localization. IEEE Access 6, 43392–43403. http://dx.doi.org/10.1109/ACCESS.2018.2856936.

Khomenko, V., Shyshkov, O., Radyvonenko, O., Bokhan, K., 2016. Accelerating recurrent neural network training using sequence bucketing and multi-GPU data parallelization. In: 1st IEEE International Conference on Data Stream Mining Processing. DSMP, pp. 100–103.

Ko, T., 2011. A survey on behaviour analysis in video surveillance applications. In: Recent Developments in Video Surveillance. InTech, pp. 279—294.

Lamb, R., Annetta, L., Hoston, D., Shapiro, M., Matthews, B., 2018. Examining human behavior in video games: The development of a computational model to measure aggression. Soc. Neurosci. 13 (3), 301–317.

Li, X., Zhang, C., Zhang, D., 2010. Abandoned objects detection using double illumination invariant foreground masks. In: 20th International Conference on Pattern Recognition. IEEE Computer Society, pp. 436–439.

Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C., 2019. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. IEEE Trans. Pattern Anal. Mach. Intell. http://dx.doi.org/10.1109/TPAMI.2019.2916873.

Mahmood Rajpoot, Q., Jensen, C.D., 2015. Video surveillance: Privacy issues and legal compliance. In: Kumar, V., Svensson, J. (Eds.), Promoting Social Change and Democracy Through Information Technology. IGI Global.

Matthews, B., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. 405 (2), 442–451.

Michailidis, M., 2017. Investigating Machine Learning Methods in Recommender Systems (Ph.D. thesis). University College London.

Micheloni, C., Canazza, S., Foresti, G., 2009. Audio-video biometric recognition for non-collaborative access granting. J. Vis. Lang. Comput. 20 (6), 353–367.

Nagai, K., Sakabe, H., Ohka, M., 2017. Finger direction recognition toward human-and-robot cooperative tasks. In: International Symposium on Micro-NanoMechatronics and Human Science. MHS, IEEE Computer Society, pp. 1–3.

Oluwatoyin, P.P., Wang, K., 2012. Video-based abnormal human behavior recognition - a review. IEEE Trans. Syst. Man Cybern. C 42 (6), 865–878.

Pan, H., He, X., Tang, S., Meng, F., 2018. An improved bearing fault diagnosis method using one-dimensional CNN and LSTM. J. Mech. Eng. 64 (8), 443–452.

Pang, K., Chen, G., Teng, W., 2013. Discovering unusual behavior patterns from motion data. In: Proceedings of the IEEE International Conference on Consumer Electronics, pp. 242–243.

Piza, E., 2018. The crime prevention effect of CCTV in public places: A propensity score analysis. J. Crime Justice 41 (1), 14–30.

Porikli, F., Ivanov, Y., Haga, T., 2008. Robust abandoned object detection using dual foregrounds. EURASIP J. Adv. Signal Process.

Raffel, C., Ellis, D.P.W., 2015. Feed-forward networks with attention can solve some long-term memory problems. CoRR abs/1512.08756.

Schuster, M., Paliwal, K., 1997. Bidirectional recurrent neural networks. IEEE Trans. Signal Process. 45 (11), 2673–2681.

Shahroudy, A., Liu, J., Ng, T.T., Wang, G., 2016. NTU RGB+D: A large scale dataset for 3D human activity analysis. In: IEEE Conference on Computer Vision and Pattern Recognition.

Sharma, S., Kiros, R., Salakhutdinov, R., 2015. Action recognition using visual attention. CoRR abs/1511.04119. URL: http://arxiv.org/abs/1511.04119.

Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.k., Woo, W.c., 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems, Vol. 28, pp. 802–810.

Simonyan, K., Zisserman, A., 2014a. Two-stream convolutional networks for action recognition in videos. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1. NIPS'14, MIT Press, Cambridge, MA, USA, pp. 568–576, URL: http://dl.acm.org/citation.cfm?id=2968826.2968890.

Simonyan, K., Zisserman, A., 2014b. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556. URL: http://arxiv.org/abs/1409.1556.

Singh, J.P., Jain, S., Arora, S., Singh, U.P., 2018. Vision-based gait recognition: A survey. IEEE Access 6, 70497–70527.

Sochor, J., Spanhel, J., Herout, A., 2019. Boxcars: Improving fine-grained recognition of vehicles using 3-D bounding boxes in traffic surveillance. IEEE Trans. Intell. Transp. Syst. 20 (1), 97–108.

Song, S., Lan, C., Xing, J., Zeng, W., Liu, J., 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: AAAI Conference on Artificial Intelligence, pp. 4263–4270.

Song, D., Tharmarasa, R., Kirubarajan, T., Fernando, X.N., 2018. Multi-vehicle tracking with road maps and car-following models. IEEE Trans. Intell. Transp. Syst. 19 (5), 1375–1386.

Stutzer, A., Zehnder, M., 2013. Is camera surveillance an effective measure of counterterrorism? Def. Peace Econ. 24 (1), 1–14.

Tian, Y., Feris, R.S., Liu, H., Hampapur, A., Sun, M.T., 2011. Robust detection of abandoned and removed objects in complex surveillance videos. IEEE Trans. Syst. Man Cybern. C 41 (5), 565–576.

Ullah, A., Muhammad, K., Del Ser, J., Baik, S.W., Albuquerque, V., 2018. Activity recognition using temporal optical flow convolutional features and multi-layer LSTM. IEEE Trans. Ind. Electron.

Vinayaga-Sureshkanth, N., Maiti, A., Jadliwala, M., Crager, K., He, J., Rathore, H., 2018. A practical framework for preventing distracted pedestrian-related incidents using wrist wearables. IEEE Access 6, 78016–78030. http://dx.doi.org/10.1109/ACCESS.2018.2884669.

Wang, X., 2013. Intelligent multi-camera video surveillance: A review. Pattern Recognit. Lett. 34 (1), 3–19.

Wang, X., Girshick, R.B., Gupta, A., He, K., 2017. Non-local neural networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7794–7803.

Wen, J., Gong, H., Zhang, X., Hu, W., 2009. Generative model for abandoned object detection. In: 16th IEEE International Conference on Image Processing. ICIP, IEEE Computer Society, pp. 853–856.

Yeung, S., Russakovsky, O., Jin, N., Andriluka, M., Mori, G., Li, F., 2015. Every moment counts: Dense detailed labeling of actions in complex videos. CoRR abs/1507.05738. URL: http://arxiv.org/abs/1507.05738.

Yu, T.H., Moon, Y.S., 2009. Unsupervised real-time unusual behavior detection for biometric-assisted visual surveillance. In: Proceedings of the Third International Conference on Advances in Biometrics. Springer-Verlag, Berlin, Heidelberg, pp. 1019–1029.

Yu, S., Qin, L., 2018. Human activity recognition with smartphone inertial sensors using bidir-LSTM networks. In: 3rd International Conference on Mechanical, Control and Computer Engineering. ICMCCE, IEEE Computer Society, pp. 219–224.

Yun, K., Honorio, J., Chattopadhyay, D., Berg, T.L., Samaras, D., 2012. Two-person interaction detection using body-pose features and multiple instance learning. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on. IEEE.