PRIVACY-PRESERVING ONLINE HUMAN BEHAVIOUR ANOMALY DETECTION BASED ON BODY MOVEMENTS AND OBJECTS POSITIONS

Federico Angelini *Jiawei Yan* *Syed Mohsen Naqvi*

Intelligent Sensing and Communications Research Group, Newcastle University, UK

ABSTRACT

Human behaviour anomaly detection is crucial for modern artificial intelligence systems. However, privacy protection plays a great role in the realization. In this paper, an online privacy-preserving anomaly detector is presented. The proposed method is able to discriminate on human subject body movements, postures and interactions with the surrounding objects, preserving subject privacy in all the tuning, training and testing stages. ActionXPose, Single Shot MultiBox Detector and Support Vector Machine are exploited for the proposed semi-supervised anomaly detector. The method successfully detects abnormal human behaviours, including unexpected body movements and misplaced objects. A new dataset ISLD-A is also proposed[1], providing suitable benchmark for performance evaluation[2].

***Index Terms*—** Privacy Protection, Anomaly Detection, ActionXPose, SSD, Human Behaviour

1. INTRODUCTION

Anomaly Detection (AD) is one of the key tasks for a modern automated surveillance system, with huge impact in many field such as healthcare, sport analysis and security in public places [1]. Video based approaches often fully rely on colour information, raising important privacy concerns, especially, for healthcare and security applications. In this paper, a novel privacy-preserving method for complex human behaviour AD is presented.

One possible option to preserve target privacy in Human Action Recognition (HAR) and AD is by using Kinect [2]. Despite its accredited HAR performance, it suffers from outdoor light conditions and it works only in limited range scenarios (up to 5-6 meters) [3]. Moreover, Kinect installation considerably raises setup costs.

Another possible solution is by leveraging Low Resolution (LR) video data. For example, authors in [4] presented interesting multi-camera results for the HAR task. However, overlooking the limited range of studied human actions, results show that pixels resolution plays a great role in HAR, followed by the number of available cameras in overlapping field of view and frame rate. To compensate performance degradation due to LR data, multi-camera coverage can help at the price of higher installation costs and synchronization issues.

A completely different approach relies on WiFi based action recognition. Authors in [5] shown HAR exploiting WiFi signals reflections. Despite the novelty of the work, authors highlighted limitations regarding the location-dependency training. Moreover, it is still not clear how much WiFi reflected signals can effectively differentiate human actions in complex environments where multiple targets are present. Moreover, it requires good area coverage with dedicated devices, which implies again important installation constrains.

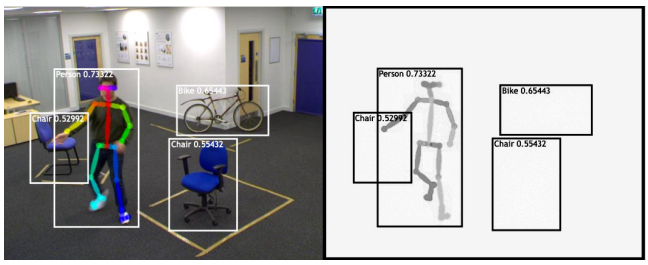


Fig. 1. (Left) RGB data, showing significant identity information, is pre-processed by OpenPose, ActionXPose and SSD. (Right). Privacy-preserving data processed by the proposed method.

On the other hand, privacy issues can be mitigated by considering tracking based methods [6, 7, 8, 9]. However, they discard all information about human action [10, 11, 12]. Background subtraction based methods are also a possible solution [13]. However, cluttered backgrounds, occlusions and self-occlusions are challenges for these methods [14]. Last but not least, Convolutional Neural Network (CNN) based features for AD [15] exploits privacy-related data only at the early stage of the processing. However, extracted features are not designed as meaningful representation for human operators [16], having been defined by a neural network optimisation, compromising further system tuning in the absence of colour data. Thus, also in this case, privacy-protection cannot be guaranteed.

Considering all above mentioned limitations, we propose a joint approach using our ActionXPose [17] and Single Shot MultiBox Detector (SSD) [18] for semi-supervised [19] anomaly detection. ActionXPose is a HAR model which relies on OpenPose [20] to discard colour data by extracting body poses. SSD is an object detector which provides only objects labels and bounding boxes. The proposed approach is online and require neither multiple camera installation nor colour data storage and also preserves target privacy. Furthermore, later-stage tuning from human operators is still possible without storing privacy-related data by using user-friendly data representation (Fig. 1).

To the best of our knowledge, none of the publicly available datasets are suitable for the considered AD problem. For example, in the popular Avenue dataset [21], many activities that are *normal* in terms of body movements, for this dataset rationale they represent *abnormal* events. Regarding the popular UCSD dataset [22], abnormal events consist mostly of trucks and cyclists in the scene, while we are interested in focusing on more complex human actions such as *picking-up*, *moving-chair*, *jumping-jacks*, *boxing*, *sitting* without loosing the attention to the human-objects interaction. Therefore, we

[1]ISLD-A will be available at http://www.intellsensing.com/research/information-processing/multimodal-surveillance/

[2]Demo is available at https://youtu.be/VJD9tYPzHPQ

propose a more suitable ISLD-A dataset for performance evaluation.

Thus, our contribution is twofold:

- a new privacy-preserving online semi-supervised algorithm for AD, which is able to specifically detect complex human behaviour anomalies, including human-objects interactions;
- a new dataset, ISLD-A, designed for complex human behaviour anomaly detection, to cover the limitations of publicly available datasets in the field.

The proposed method for AD is particularly useful for scenerios where high privacy-protection is required, such as human *fall* detection for assisted living, *gate* monitoring and people *fight* detection for security applications.

2. PRELIMINARIES

Our ActionXPose is a real-time *posture-level* Human Action Recognition (HAR) algorithm [17]. Posture-level refers to the fact that ActionXPose only uses body poses provided by OpenPose [20] (Fig. 1) to perform HAR. RGB colour information is only exploited to extract body poses at the early stage and discarded. ActionXPose exploits Multivariate Long Short-Term Memory (MLSTM-FCN) network [23] to model body pose patterns, in combination with Self-Organising Map (SOM) for unsupervised clustering [24].

In this paper, ActionXPose has been pre-trained on 18 human actions using MPOSE and ISLD datasets [17], alongside with new video clips recorded in our state-of-the-art lab to improve viewpoint changes robustness. As opposite to [17], short-time-clips (30 frames) were processed during the training phase. This time restriction allows ActionXPose to deal with quick human motion.

Regarding objects detection, SSD is a state-of-the-art real-time CNN based method for this task [18]. It is able to provide, frame-by-frame, object class labels and bounding boxes for all detected objects (Fig 1). In this paper, SSD has been pre-trained for the detection of 20 objects by using Pascal dataset [25]. Depending on the scenario, transfer learning [26] can be applied to tune SSD in order to fulfil the required recognition abilities.

3. ANOMALY DETECTION

3.1. Problem Statement

In this paper, we consider two types of possible anomalies, respectively Body-Movements (BM) related anomalies and Object-Position (OP) related anomalies. In other words, we propose an AD algorithm able to detect those anomalies due to unexpected body movements and unusual objects positions. Moreover, for simplicity, we focus on recordings where only one target is visible at a time. However, generalisation to multiple-targets are straightforward due to the exploited detector. Thus, we propose a semi-supervised anomaly detection framework to jointly detect BM and OP anomalies. *Normal data* consists of video clips containing expected human behaviours only. As opposite, *test data* consists of video clips where normal and abnormal behaviours can occur. Hence, the proposed method aims to learn a behaviour patterns model from normal data in order to estimate normal/abnormal responses for testing data.

Let $\mathbb{D} = \mathbb{N} \cup \mathbb{T}$ a single target dataset, where $\mathbb{N}$ represents normal data while $\mathbb{T}$ represents testing data. In particular, let $\mathbb{N} = \{\mathbf{s}_j\}_{j=1}^{n}$ be the normal video data subset containing n clips of arbitrary time length, where $\mathbf{s}_i$ represents the j-th RGB video clip. Let $\mathbb{T} = \{\mathbf{v}_i, g_{\mathbf{v}_i}(t)\}_{i=1}^{m}$ be the video testing dataset containing m clips, where $\mathbf{v}_i$ represents the i-th RGB video clip and $g_{\mathbf{v}_i}(t)$ represents the ground truth for $\mathbf{v}_j$. In particular, $g_{\mathbf{v}_j}(t)$ is a function that associates each frame t of $\mathbf{v}_j$ with a binary response, i.e. normal/abnormal label, and it can be defined as:

$$g_{\mathbf{v}_j}(t) = \begin{cases} 0 & \text{if } t \text{ is labelled as normal} \\ 1 & \text{if } t \text{ is labelled as abnormal} \end{cases} \quad (1)$$

Since $\mathbb{D}$ is a single target dataset, anomaly localisation is provided by the target detector. In the case of multiple-targets, a tracking mechanism, such as our recent tracking system [6], can be exploited to preserve the association between targets identities and detected anomalies across frames.

Thus, the goal of this paper is to propose a strategy to compute $G_{\mathbf{v}_i}(t) \approx g_{\mathbf{v}_i}(t)$ for all $i = 1, \ldots, m$, i.e. to effectively estimate whether a testing frame in $\mathbb{T}$ is normal or abnormal, analysing the target behaviour and comparing it with the behaviour patterns learnt with normal data $\mathbb{N}$.

3.2. Body Motion Based Anomaly Detection

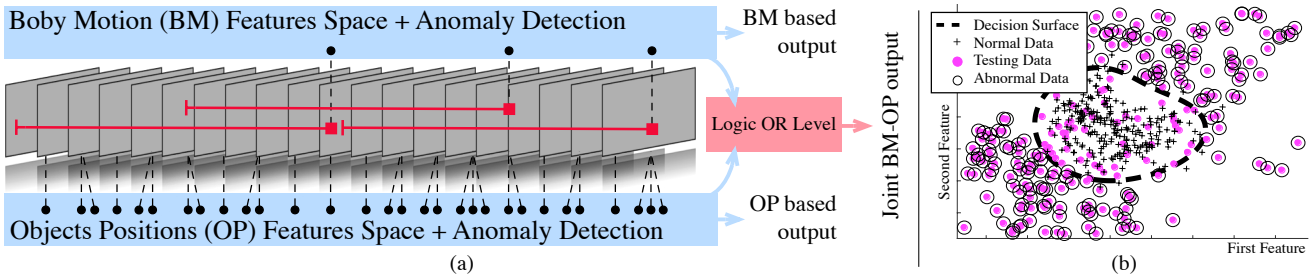
Body Motion (BM) based AD is only exploiting human body motion information. The idea is to collect features from expected body movements provided by normal data. Thus, a model for classifying a testing movement as normal or abnormal is trained. This paper proposes to use ActionXPose as features extraction method for BM analysis. In fact, despite ActionXPose has been pre-trained for classification purposes, by simply removing the classification SVM layer in the MLSTM-FCN network embedded into ActionXPose, we achieve a feature extraction algorithm for short-time body motions that does not need to be further trained for the AD task. Despite the nature of the ActionXPose training classes, not only normal/abnormal motion detection regarding these classes can be performed but also regarding new actions and movements, such as the human *falling* action.

Overlapping short-time-clips from normal data $\mathbb{N}$ are provided to ActionXPose which in turn extracts features vectors $\mathbf{b} \in \mathbb{R}^{136}$ (Fig. 2-(a)). Thus, due to the short-time-clips overlapping, each feature vector $\mathbf{b}$ is associated with a batch of 30 frames $\{t_1, \ldots, t_{30}\}_{\mathbf{b}}$, where some frames can be associated with multiple feature vectors. Thus, for each frame t, $I(t) = \{\mathbf{b}_h\}_h$ is the set of all possible feature vectors that are associated with t.

Collected $\mathbf{b}$ vectors from normal data are subsequently embedded into a common multidimensional Cartesian space. Thus, one-class SVM classifier is trained for semi-supervised AD. This approach has been extensively used for AD when only normal data is available and it represents a standard approach for such problems [27]. The idea is that only single class training data is provided to the classifier (normal behaviour data). Thus, the goal is to define a region of the features space where normal feature vectors occur with high probability. The *kernel trick* allows the region to be possibly close and nonlinear [27], depending on the normal data space distribution. In Fig. 2-(b), a 2D example is shown to describe such one-class SVM classifier, where normal data have been used to train the AD model, defining the decision surface. Thus, in the model testing phase, testing feature vectors which lay outside the decision surface have been classified as abnormal.

Therefore, for each testing feature vector $\mathbf{b}$, we can define a *response* $R(\mathbf{b})$ as follows:

$$R(\mathbf{b}) = \begin{cases} 0 & \text{if } \mathbf{b} \text{ is labelled as normal} \\ 1 & \text{if } \mathbf{b} \text{ is labelled as abnormal} \end{cases} \quad (2)$$



(a) (b)

Fig. 2. (a) *Proposed AD strategy*. The video is split into short-time-clips (red line) 30 frames long with 15 overlapping frames, to allow ActionXPose to extract features from each short-time-clip. Thus, obtained features are embedded into a common space (BM Features Space), where only features from normal data are used to train the one-class SVM classifier for anomaly detection. Therefore, by using the trained SVM model, testing features are classified as *normal* or *abnormal*. Conversely, OP related features are extracted from *each frame* by using SSD and embedded into a common space (OP Features Space) for anomaly detection training and testing. OP based normal/abnormal responses which constitute OP based decision. The additional Logic OR level is used to combine these two outputs for an effective joint BM-OP AD. (b) *Example of bi-dimensional one-class SVM features space*. Training data (black crosses) are used to train the SVM model. Thus, testing data (magenta dots) are classified as *normal* or *abnormal* (black circles). The boundary decision surface (cyan line) is given by the SVM training phase.

Hence, the BM feature space provides only BM related output $G^{BM}_{\mathbf{v}_i}(t)$, which is a binary function associating each frame with a *normal/abnormal* classification depending on the correspondent feature vector. It can be defined as:

$$G^{BM}_{\mathbf{v}_i}(t) = \begin{cases} 0 & \text{if} \quad R(\mathbf{b}) = 0, \quad \forall \mathbf{b} \in I(t) \\ 1 & \text{if} \quad \exists\, \mathbf{b} \in I(t) \;\; s.t. \;\; R(\mathbf{b}) = 1 \end{cases} \tag{3}$$

which represents the BM based approximation of $g_{\mathbf{v}_i}(t)$.

3.3. Object Position Based Anomaly Detection

Object Position (OP) based AD is only relying on SSD bounding boxes of detected objects. We propose to scan $\mathbb{D}$ with SSD frame-by-frame. Thus, object centroids coordinates are computed from bounding boxes. As already explained in Section 3.2, the idea is to learn from $\mathbb{N}$ with a SVM model the normal objects positions, including the human target position.

Centroids coordinates not only contain information about objects locations, but also about mutual Euclidean distances between them. This is particularly useful in scenarios where some key objects cannot be misplaced by the human target. Other example can be the one where vehicles and pedestrian are expected to drive and walk in certain areas and automatic detection of anomalies is required.

Thus, as shown in Fig. 2-(a), OP based AD consists of extracting objects positions as:

$$\{(x,y)_{k,1}, \ldots, (x,y)_{k,J_k}\}_{k=1}^{K} \tag{4}$$

where J_k represents the number of occurrences of the k-th object. For example, if $k = 1$ represents *chair*, multiple chairs can be present in the scene, i.e. $J_1 = 1, 2, 3, \ldots$. Thus, for each single occurrence of each detected object, we can define *feature vectors* as:

$$\begin{aligned} &\mathbf{q} = [x_{1,j_1}, y_{1,j_1}, \ldots, x_{K,j_K}, y_{K,j_K}] \in \mathbb{R}^{2K}, \\ &\forall j_1 \in [1, \ldots, J_1], \ldots, \forall j_K \in [1, \ldots, J_K] \end{aligned} \tag{5}$$

where $\mathbf{q}$ is defined as the concatenation of x and y entries of each object occurrence. This feature vector is short and captures relevant information about OP, despite the presence of multiple objects of the same type in the scene. Therefore, since multiple objects can be present at the same time, multiple feature vectors can be extracted from each frame. When object data is missing, it can be estimated using the previous available frame.

Feature vectors can be embedded into a common features space. Thus, features from $\mathbb{N}$ can be used to train a one-class SVM classifier for AD. In this case, although multiple feature vectors are available for each frame, we can still define the approximation $G^{OP}_{\mathbf{v}_i}(t)$ as follows

$$G^{OP}_{\mathbf{v}_i}(t) = \begin{cases} 0 & \text{if} \quad R(\mathbf{q}) = 0 \quad \forall \mathbf{q} \in I(t) \\ 1 & \text{if} \quad \exists\, \mathbf{q} \in I(t) \;\; s.t. \;\; R(\mathbf{q}) = 1 \end{cases} \tag{6}$$

where $I(t)$ represents the set of feature vector associated with t, $\mathbf{q}$ as in (5) and $R(\mathbf{q})$ is the response for vector $\mathbf{q}$ as defined in (2).

3.4. Joint BM-OP Anomaly Detection

In this section, a simple combination rule is proposed in order to combine BM and OP outputs, namely $G^{BM}_{\mathbf{v}_i}(t)$ and $G^{OP}_{\mathbf{v}_i}(t)$ for a joint BM-OP AD. Let assume that BM and OP abnormal events are theoretically *independent*. Thus, given a probability distribution $\mathbb{P}(\cdot,\cdot)$ for joint abnormal BM-OP based events, we can write

$$\begin{aligned} &\mathbb{P}\left(\{g_{\mathbf{v}_i}(t) = 1 \mid BM\}, \{g_{\mathbf{v}_i}(t) = 1 \mid OP\}\right) = \\ &= \mathbb{P}(g_{\mathbf{v}_i}(t) = 1 \mid BM) \;\times\; \mathbb{P}(g_{\mathbf{v}_i}(t) = 1 \mid OP) \approx \\ &\mathbb{P}(\mathbf{b} \in I(t) \;\; s.t. \;\; R(\mathbf{b}) = 1 \mid \text{SVM}_{\text{BM}}) \times \\ &\times\, \mathbb{P}(\mathbf{q} \in I(t) \;\; s.t. \;\; R(\mathbf{q}) = 1 \mid \text{SVM}_{\text{OP}}) \end{aligned} \tag{7}$$

where SVM_{BM} and SVM_{OP} represent respectively the BM based SVM anomaly detection model and the OP based SVM one, defined in Sections 3.2 & 3.3. Thus, we propose to approximate $g_{\mathbf{v}_i}$ with $G_{\mathbf{v}_i}$ such that

$$G_{\mathbf{v}_i}(t) = \begin{cases} 0 & \text{if} \quad G^{BM}_{\mathbf{v}_i}(t) = 0 \;\wedge\; G^{BM}_{\mathbf{v}_i}(t) = 0 \\ 1 & \text{otherwise} \end{cases} \tag{8}$$

Such definition correspond to impose *Logic OR* condition for anomaly detection on BM and OP outputs. Thus, (8) is called *Logic OR Level* (Fig. 2-(a)).

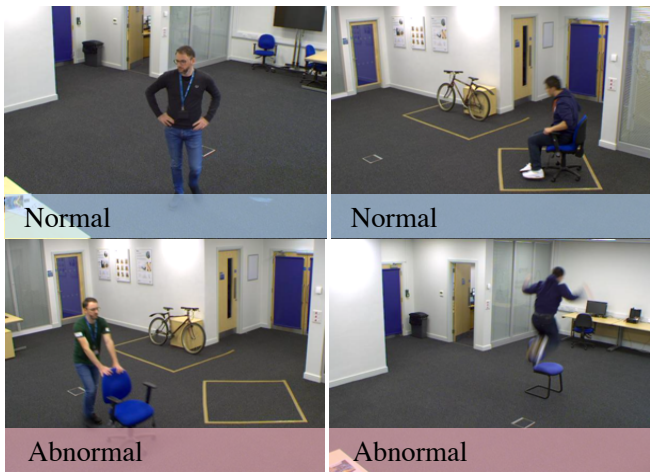


Fig. 3. Proposed dataset example frames, showing normal/abnormal body movements and expected objects positioning.

4. EXPERIMENTS

In order to evaluate the proposed method, we recorded a dataset, namely ISLD-A, in the Intelligent Sensing Lab, which will be public soon. Video clips are recorded from two static cameras, showing two different subjects performing spontaneous behaviours involving two objects, namely *chair* and *bike*. In Fig. 3, some example frames of the proposed dataset are depicted. This dataset includes three sub-datasets for different evaluations, namely:

- BM based dataset $\{\mathbb{N}_1, \mathbb{T}_1\}$ (BMbD): we asked the subjects to be spontaneous in the scene, performing only those actions they felt *reasonably acceptable* in a working environment. Resulting clips have been added to $\mathbb{N}_1$ dataset, i.e. normal dataset. Subsequently, we left them free to perform any actions they wanted and we added these clips into the $\mathbb{T}_1$ dataset;
- OP based dataset $\{\mathbb{N}_2, \mathbb{T}_2\}$ (OPbD): normal data $\mathbb{N}_2$ is constituted by clips where two objects, namely *chair* and *bike*, have been constrained to remain in a pre-defined area of the lab while the subject is free to move in all locations. Conversely, testing data $\mathbb{T}_2$ is made by clips where the subjects, occasionally, move the objects out from the allowed areas;
- Joint BM-OP based dataset $\{\mathbb{N}_3, \mathbb{T}_3\}$ (JBMOPbD): in this case, $\mathbb{N}_3$ include video clips in $\mathbb{N}_1$ alongside with new recordings to impose the objects positions constrains. Thus, in sequences $\mathbb{T}_3$, the subjects were asked to perform freely any behaviour, either with or without moving objects.

By comparing $\mathbb{N}_i$ and $\mathbb{T}_i$ for $i = 1, 2$ and 3, ground truth for testing clips were set by a human operator. Standard performance metrics, i.e. accuracy, sensitivity and specificity, are reported for performance evaluation [11]. In Table 1, we provide results of BM based and OP based AD. Additionally, we also provide results for Joint BM-OP AD. In particular, for the Joint BM-OP case, we reported the result for the single modality cases, in order to emphasise the advantages of the Joint BM-OP approach over the single BM and OP anomaly detection. Since ActionXPose and SSD have been pre-trained on different datasets for classification purposes, no further training is needed to complete the AD task. Thus, only the one-class SVM models must be trained. We report averaged results over 10 different trainings for the one-class SVM models, including standard deviations.

BMdD, $\mathcal{T}(\mathbb{N}_1) = 21$m:34s, $\mathcal{T}(\mathbb{T}_1) = 19$m:42s			
Modality	Accuracy (%)	Sensitivity (%)	Specificity (%)
BM	81.94±0.19	81.55±1.59	83.79±0.95
OPbD, $\mathcal{T}(\mathbb{N}_2) = 3$m:12s, $\mathcal{T}(\mathbb{T}_2) = 5$m:30s			
Modality	Accuracy (%)	Sensitivity (%)	Specificity (%)
OP	89.53±2.69	93.73±0.69	86.59±4.96
JBMOPbD, $\mathcal{T}(\mathbb{N}_3) = 21$m:38s, $\mathcal{T}(\mathbb{T}_3) = 12$m:44s			
Modality	Accuracy (%)	Sensitivity (%)	Specificity (%)
BM	68.94± 0.01	61.15± 0.02	78.86± 0.01
OP	81.80± 0.23	72.00±1.52	87.34±1.59
J. BM-OP	86.01±0.19	91.27±1.20	79.94±1.45

Table 1. Results for ISLD-A dataset. $\mathcal{T}(\circ)$ is the total time length. Accuracy and precision results are given on average for 10 different SVM model training, including obtained standard deviation σ.

Results for BMdB and OPbD show the effectiveness in terms of accuracy of the BM and OP based approach in different contexts. Moreover, in the JBMOPbD, Joint BM-OP based AD accuracy outperforms BM and OP based approach accuracies. From a heuristic approach, viewing the output videos, it turns out that the proposed methods is able to detect almost the 100% of the abnormal sub-sequences. The major errors source, which drops the accuracy results, is the sub-optimal localisation of the starting and ending frames for each abnormal sequence.

It is worth mentioning that some video clips contain human actions which are not part of the ActionXPose pre-training, such as *falling* and other body motions. However, the proposed method can effectively detect them as anomalies. Such general ability in detecting anomalies, despite the limited range of class actions ActionXPose has been trained on (18 classes), makes this method particularly useful for example for human *fall detection*. In fact, all *falling* action sequences in the testing data have been successfully detected as anomalies.

Regarding processing time, both the body pose detector and the objects detector are the *bottle-neck* of the system. However, they can both run with real-time performance. To justify online performance, we remark that, as mentioned in Section 3, for the decision at frame t only the current and the previous 30 frames are taken into account.

5. CONCLUSIONS

The proposed method is able to successfully detect abnormal events in the tested videos, consisting of abnormal human body behaviour as well as unexpected interaction with objects, while effectively preserving target privacy. The proposed processing is online and shows detection abilities for a wide range of abnormal human events, including for example *falling* action. Future work will be on improving the abnormality time localisation and on implementing targets tracking for multiple-targets anomaly detection. Moreover, advanced versions of SSD can be pre-trained, in order to extend object detection abilities. Furthermore, since the core of the proposed AD method is ActionXPose, further improvements on this side will be also considered.

6. REFERENCES

[1] C. Savitha and D. Ramesh, "Motion detection in video surviellance: A systematic survey," in *IEEE 2nd International Conference on Inventive Systems and Control (ICISC)*, 2018, pp. 51–54.

[2] Pichao Wang and Philip O Ogunbona, "RGB-D-based Motion Recognition with Deep Learning: A Survey," *IJCV*, vol. TBA, no. June, pp. 1–34, 2017.

[3] Benjamin Langmann, Klaus Hartmann, and Otmar Loffeld, "Depth Camera Technology Comparison and Performance Evaluation," *1st International Conference on Pattern Recognition Applications and Methods*, vol. 2, pp. 438–444, 2012.

[4] Ji Dai, Jonathan Wu, Behrouz Saghafi, Janusz Konrad, and Prakash Ishwar, "Towards privacy-preserving activity recognition using extremely low temporal and spatial resolution cameras," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 68–76.

[5] Jen-Yin Chang, Kuan-Ying Lee, Kate Ching-Ju Lin, and Winston Hsu, "WiFi action recognition via vision-based methods," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2782–2786.

[6] Zeyu Fu, Pengming Feng, Federico Angelini, Jonathon Chambers, and Syed Mohsen Naqvi, "Particle PHD Filter Based Multiple Human Tracking Using Online Group-Structured Dictionary Learning," *IEEE Access*, vol. 6, pp. 14764–14778, 2018.

[7] Pengming Feng, Wenwu Wang, Satnam Dlay, Syed Mohsen Naqvi, and Jonathon Chambers, "Social Force Model-Based MCMC-OCSVM Particle PHD Filter for Multiple Human Tracking," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 725–739, 2017.

[8] Ata Ur-Rehman, Syed Mohsen Naqvi, Lyudmila Mihaylova, and Jonathon A. Chambers, "Multi-Target Tracking and Occlusion Handling With Learned Variational Bayesian Clusters and a Social Force Model," *IEEE Transactions on Signal Processing*, vol. 64, no. 5, pp. 1320–1335, 2016.

[9] Pengming Feng, Wenwu Wang, Syed Mohsen Naqvi, and Jonathon Chambers, "Adaptive Retrodiction Particle PHD Filter for Multiple Human Tracking," *IEEE Signal Processing Letters*, vol. 23, no. 11, pp. 1592–1596, 2016.

[10] Karuna B. Ovhal, Sonal S. Patange, Reshma S. Shinde, Vaishnavi K. Tarange, and Vijay A. Kotkar, "Analysis of anomaly detection techniques in video surveillance," in *International Conference on Intelligent Sustainable Systems (ICISS)*, 2017, pp. 596–601.

[11] Avinash Ratre and Vinod Pankajakshan, "Tucker tensor decomposition-based tracking and Gaussian mixture model for anomaly localisation and detection in surveillance videos," *IET Computer Vision*, vol. 12, no. 6, pp. 933–940, 2018.

[12] Shyma Zaidi, B Jagadeesh, K V Sudheesh, and Arlene A Audre, "Video Anomaly Detection and Classification for Human Activity Recognition," in *International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*. 2017, pp. 544–548, IEEE.

[13] Federico Angelini, Zeyu Fu, Sergio A. Velastin, Jonathon A. Chambers, and Syed Mohsen Naqvi, "3D-Hog Embedding Frameworks for Single and Multi-Viewpoints Action Recognition Based on Human Silhouettes," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4219–4223.

[14] Akash Gandhamal and Sanjay Talbar, "Evaluation of background subtraction algorithms for object extraction," in *IEEE International Conference on Pervasive Computing (ICPC)*, 2015, pp. 1–6.

[15] Ryota Hinami, Tao Mei, and Shin'ichi Satoh, "Joint Detection and Recounting of Abnormal Events by Learning Deep Generic Knowledge," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3639–3647.

[16] Yann A. LeCun, Yoshua Bengio, and Geoffrey E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[17] Federico Angelini, Zeyu Fu, Yang Long, and Syed Mohsen Naqvi, "ActionXPose: A Novel 2D Multi-view Pose-based Algorithm for Real-time Human Action Recognition," *arXiv (to be submitted to IEEE Transactions on Cybernetics)*, 2018.

[18] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "SSD: Single Shot MultiBox Detector," in *ECCV*, 2016.

[19] Markus Goldstein and Seiichi Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PLoS ONE*, vol. 11, no. 4, 2016.

[20] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," in *CVPR*, 2017, pp. 7291–7299.

[21] Cewu Lu, Jianping Shi, and Jiaya Jia, "Abnormal Event Detection at 150 FPS in MATLAB," in *IEEE International Conference on Computer Vision*, 2013, pp. 2720–2727.

[22] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos, "Anomaly detection in crowded scenes," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1975–1981.

[23] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Samuel Harford, "Multivariate LSTM-FCNs for Time Series Classification," *arXiv*, 2018.

[24] Teuvo Kohonen, *Self-organizing maps*, Springer, Berlin, 3rd edition, 2001.

[25] M Everingham, L Van˜Gool, C K I Williams, J Winn, and A Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[26] Ian Goodfellow, *Deep learning*, The MIT Press, 2016.

[27] Colin Campbell, *Learning with support vector machines*, Morgan & Claypool, 2011.