RGB-D Fall Detection via Deep Residual Convolutional LSTM Networks

Ahmed Abobakr*, Mohammed Hossny*, Hala Abdelkader† and Saeid Nahavandi*
*Institute for Intelligent Systems Research and Innovation (IISRI)
Deakin University, Australia
Email: aabobakr@deakin.edu.au
†Faculty of Computers and Information, Cairo University, Egypt
Email: h.abdelkader@fci-cu.edu.eg

***Abstract*—The development of smart healthcare environments has witnessed impressive advancements exploiting the recent technological capabilities. Since falls are considered a major health concern especially among older adults, low-cost fall detection systems have become an indispensable component in these environments. This paper proposes an integrable, privacy preserving and efficient fall detection system from depth images acquired using a Kinect RGB-D sensor. The proposed system uses an end-to-end deep learning architecture composed of convolutional and recurrent neural networks to detect fall events. The deep convolutional network (ConvNet) analyses the human body and extracts visual features from input sequence frames. Fall events are detected via modeling complex temporal dependencies between subsequent frame features using Long-Shot-Term-Memory (LSTM) recurrent neural networks. Both models are combined and jointly trained in an end-to-end ConvLSTM architecture. This allows the model to learn visual representations and complex temporal dynamics of fall motions simultaneously. The proposed method has been validated on the public URFD fall detection dataset and compared with different approaches, including accelerometer based methods. We achieved a near unity sensitivity and specificity rates in detecting fall events.**

***Keywords*—Kinect, RGB-D, fall detection, ConvNet, LSTM, ConvLSTM.**

I. Introduction

Falls are considered a major health concern, especially for elderly people. The World Health Organisation [1] have named fall events as the second leading cause of accidental or unintentional injury deaths. Worldwide, it has been estimated that 37 million severe falls that require medical attention occur each year [1]. A fall is defined as an event that results in a person coming to rest inadvertently on the ground or other lower level structures. Elderly people have a higher risk of having a serious fall due to biological changes of the natural ageing process [2]. Falls also may cause psychological consequences such as losing confidence and independence in addition to other long term effects. Furthermore, delayed medical assistance and staying on the floor for long time increases the risk of both physical and psychological complications [3]. Therefore, fall detection systems are actively investigated and integrated in healthcare environments [4].

A robust fall detection system can be defined as an assistive device which aims at detecting and alerting fall incidents. Hence, its main objective is to distinguish between fall events and normal activities of daily living (ADL) [4]. The significant similarities of some ADL activities to falls challenges the robustness of fall detection systems. Thus, the research and development of robust fall detectors has witnessed a dramatic increase over the past few years. Fall detection studies have been recently summarised into two main categories: wearable devices based and context-aware systems [4]. Accelerometer devices are the most commonly used wearable sensors for fall detection. Readings from the accelerometer attached to the human body are evaluated using thresholding [5] or machine learning methods [6] to detect fall events. Despite the effectiveness of wearable devices, they have limitations such as battery lifetime, being easily disconnected and forgotten [7]. Moreover, wearing electronic devices is not preferable for the ageing societies [7]–[9]. Context-aware systems, on the other hand, rely on sensors deployed in the environment such as

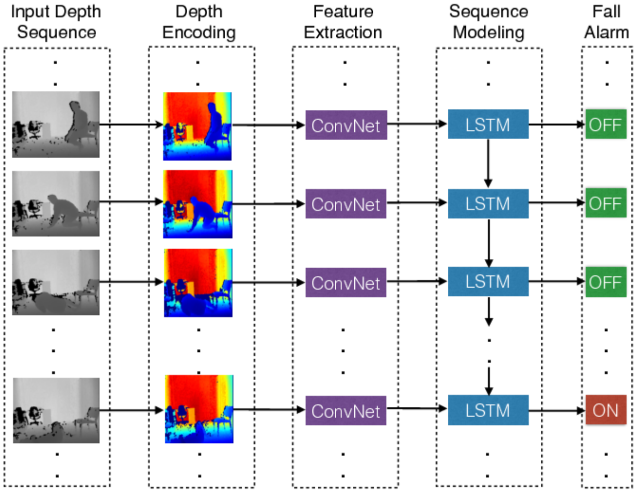

Fig. 1: The proposed fall detection method. An end-to-end ConvLSTM model has been trained to detect fall actions. First, an input video of depth frames is colorised and passed through a deep ConvNet model for extracting visual features. Second, the extracted sequence features are fed to the recurrent LSTM module to analyse temporal dynamics between frames and detect the occurrence of fall events.

978-1-5386-6602-9/18/$31.00 ©2018 IEEE

floor vibration sensors, microphones and cameras to detect falls. Relying on sound or vibration signals is challenged by frequent false alarms that may result from falling objects. Hence, the combination of cameras, computer vision and machine learning is dominating the research in fall detection [4].

This paper proposes a vision based integrable and automated fall detection system. The fall events are detected using an end-to-end deep machine learning model composed of convolutional and recurrent neural networks. The deep convolutional network (ConvNet) extracts visual features from input sequences of depth frames. We use a ConvNet architecture that follows the residual learning approach (ResNet) [10] to optimise the visual representations. The key aspect in ResNet is that, layers learn a residual function with reference to layers input instead of learning an unreferenced transformation. It features easier optimisation, computational efficiency and has high accuracy gains with deeper networks [10]. The Long-Short-Term-Memory (LSTM) recurrent neural network module is used on top of the ResNet model to learn temporal dynamics that discriminates fall and non-fall events. Figure 1 depicts an overview of the proposed method. This ConvLSTM combination is jointly trained end-to-end, and allows learning visual perceptual representations and temporal dynamics simultaneously.

The proposed system relies on RGB-D sensors for input data acquisition. Traditional RGB cameras measure color intensities with a high dependence on illumination. This imposes difficulties in modeling 3D objects and performing foreground separation [9], [11]. On the other hand, depth imaging technologies estimate the distance of 3D points in the scene away from the imaging plane. This feature has dramatically simplified a wide range of vision tasks. Depth cameras use infrared structured light [12] to acquire depth information enabling illumination independence. Further, they can work in low/no light conditions enabling privacy preservation which is a design requirement for fall detection systems [7]. The popular Microsoft Kinect RGB-D sensor has provided an accurate learning signal [13] for a wide range of visual reasoning applications such as human pose estimation [14], [15]. Further, Webster and Celik [16] demonstrated the use of Kinect in elderly care and stroke rehabilitation [17]. Although Kinect v2, the second version of Microsoft Kinect, produces less noisy depth images, we chose to constraint the proposed solution to Kinect v1. This is due to the increased power consumption and cooling requirements of Kinect v2, which is not suitable for deployment on embedded devices [18]. These requirements are dictated by the time of flight imaging technology in Kinect v2 as compared to the structured light technology in Kinect v1.

The rest of this paper is organised as follows. Section II presents a brief survey on recent fall detection methods. Section III describes the proposed fall detection method. Experiments and results are discussed in Section IV. Finally, conclusion and future work are highlighted in Section V.

II. Related Work

Enhancing smart healthcare and medical services is a multidisciplinary area of research that receives high interest. Fall detection in particular has witnessed much focus due to the physical and psychological complications of falls on elders [3]. This section reviews the most recent context-aware and vision-based fall detection methods. These approaches analyse the temporal dynamics in input videos to detect fall incidents.

The use of traditional 2D stereo camera for fall detection was demonstrated in Rougier2011 by analysing shape deformation and [19] via 3D head tracking. However, 2D approaches have much difficulties such as ambiguities of appearance, occlusions, modeling 3D objects and high dependency on illumination [9], [11]. Therefore, a growing trend is to use the depth cameras as they provide much richer geometrical information and facilitate preprocessing tasks such as background subtraction and objects delineation. Further, privacy concerns and low/no light challenges limit applications of fall detection systems relying on RGB cameras [7].

The Kinect depth sensor has been extensively used in the development of fall detection systems [7], [9], [20], [21]. Stone et al. [9] characterised person vertical state in depth frames over time. Then, a temporal segmentation was used to identify ground events. Five features were extracted for each ground event and a random decision forest (RDF) was used to compute a confidence that a fall preceded a ground event. In this method, the vertical state characterisation was done based on predefined measurements. Also, the whole body should be visible, as any segmentation artifacts of ground events will relatively affect the feature extraction process. Bian et al. [20] developed a two stage system. First, the 3D coordinates of body joints were interpolated from segmented body parts. Second, a support vector machine (SVM) classifier was used to detect the fall based on the positions of the extracted joint trajectory over time. Limitations of this method are: high dependency on head position and predefined threshold values. Abobakr et al [7], tackled these challenges via proposing a posture analysis approach that does not rely on extracting skeleton data such as head or hips positions. The articulated posture is recognised using a dense RDF model that was trained on pixel-wise feature representations. The fall event was detected using a SVM model that analyses changes in lying posture confidence overtime. However, relying on analysing temporal dynamics of posture confidence patterns makes this method sensitive to noisy depth measurements that may affect pixel votes. Also, this approach requires an expensive background rejection process and it necessitates for scene calibration whenever change occurs in scene objects configurations.

To address the aforementioned limitations, the proposed method does not rely on skeleton extraction, joints tracking, joint altitude thresholds or ground plane calibration. We propose a ConvLSTM model that combines a deep ConvNet feature extractor with a recurrent LSTM module to learn temporal dynamics of fall events. The proposed model is

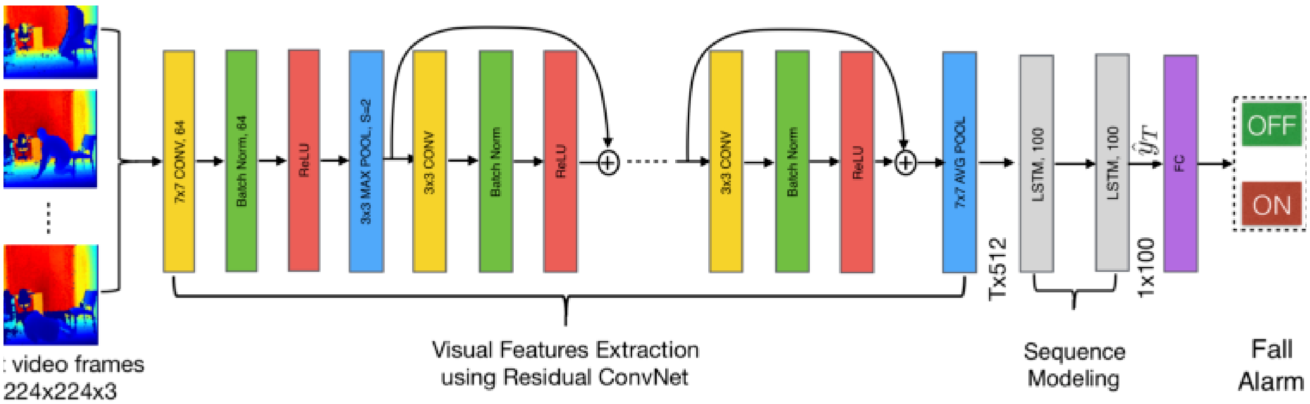

Fig. 2: Fall event detection using recurrent convolutional neural networks. An input video sequence of T depth frames is preprocessed using depth scaling, image resizing and cropping, and RGB colorisation. Then, the resulting sequence of dimensionality $T \times 224 \times 224 \times 3$ is fed to the trained residual ConvNet model to extract visual feature vectors of dimensionality $T \times 512$. Two stacked LSTM layers of 100 hidden units transform the sequence into hidden state vectors, from which we fed the last one to the fully connected layer to decide the occurrence of a fall event.

spatio-temporal and end-to-end trainable using backpropagation. The ConvLSTM model has powerful capabilities to focus on human body movements and ignore static scene objects. Therefore, it does not require background modeling and subtractions stages.

III. Proposed Method

We propose an end-to-end ConvLSTM model for fall event detection. This model combines a deep residual convolutional network ResNet with recurrent LSTM neural network module. We use a ConvNet architecture that follows the residual learning approach (ResNet) [10] to learn discriminative features from articulated body postures. The motivation for this combination is twofold. First, the ResNet model has powerful capabilities to learn and extract deep hierarchical visual features from raw input images. Second, using the extracted body features, the LSTM module can learn long-term temporal dynamics that can discriminate sequential input data, e.g., fall events. Figure 2 shows in more details the proposed ConvLSTM method for fall detection.

The ConvLSTM model maps an input video sequence X $= \{(x_t, y)\}_{t=1}^{T}$ of T depth frames to a single static output $\hat{y}_T$ representing the posterior probability distribution over two classes: (fall, non-fall). The input frames are acquired using a Kinect V1 sensor with dimensionality 480×640. At a time step t, the depth frame x_t is preprocessed via resizing and RGB colorisation to $224 \times 224 \times 3$ pixels, matching the fixed input dimensionality of the ResNet model. It performs feature transformation to map the colorised image x_t into a fixed-length feature vector $f_t \in \mathbb{R}^{d=512}$. The whole input sequence is processed in parallel producing a feature vector sequence F $= \{(f_t, y)\}_{t=1}^{T}$, where F $\in \mathbb{R}^{T \times 512}$. The resulting feature sequence is passed into the recurrent LSTM module for sequence modeling and learning.

A. Preprocessing Depth Sequences

Deep learning models have demonstrated powerful capacity in learning deep hierarchical features from raw input RGB images. Several studies have investigated the capabilities of these models to learn perceptual representations from depth images [22]–[25]. It has been concluded that depth images provide weak local gradient information of objects which makes it difficult for deep learning models to generalise and biases the ConvNet model towards detecting objects silhouettes [22], [24]. Therefore, several depth encoding methods have been proposed to make efficient use of depth measurements and provide better learning signal for the deep network.

Couprie et al. [23] combined the depth map as an additional channel with the RGB image forming a RGB-D modality. This concatenation was successful in achieving good convergence for vision tasks and semantic segmentation in particular. In [24], depth representation using the HHA encoding was proposed. In this approach, depth pixels are represented using three components; height above the ground, horizontal disparity and surface normals. Learning from HHA encoded maps has demonstrated better results than using raw depth data. However, extracting the HHA features is computationally expensive [25]. Further, it may sound analogous to the traditional feature extraction paradigm, which incorporates much prior knowledge about the input [22].

RGB colorisation [22], [25] has recently become the most commonly used depth encoding technique. This method spreads depth measurements over three RGB color channels. The values of RGB color components vary according to the distance from the depth camera, and hence, provide more powerful input signal to the ConvNet model, ResNet in this work. Depth frames of an input sequence are resized and cropped to 224×224 dimensionality. Then, depth measurements are shifted, to provide depth invariance, and normalised to $(0, 1)$ range. Finally, a RGB color map is applied to produce

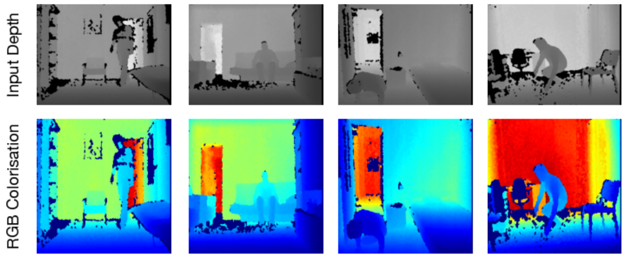

Fig. 3: Example depth preprocessing and encoding results using RGB colorisation. Input depth frames of a video sequence are resized and cropped to suit the fixed input size of the ResNet model, and their depth values are scaled to $(0, 1)$ range. Then, a jet color map is applied to produce $224\times224\times3$ RGB images.

the colorised depth image. Example preprocessing results are shown in Fig. 3.

B. Deep Features Extraction using ResNet

Extracting deep hierarchical representation of the visual world is an essential step towards building efficient visual perception systems. Deep convolutional neural networks (ConvNets) have demonstrated superior capabilities in learning such representations. ConvNets are the basic building block for the state-of-the-art methods on visual perception tasks such as object recognition [10], localisation and detection [26], and semantic segmentation [26]. The ConvNet model is a stack of computational layers; convolution (CONV), pooling or downsampling (POOL) and batch normalisation (BN), interleaved with a non-linear transformation, for instance, the rectified linear unit (ReLU). This stack of layers learns multiple levels of feature extractors with increasing level of abstraction.

Incorporating ConvNets with the residual learning paradigm has led to the ResNet architecture which ensures better and faster generalisation performance, easier optimisation and makes efficient use of network depth [10]. In this work, we use the ResNet model for learning extracting visual features from input depth videos. The ResNet model is composed of residual blocks where each block learns a residual mapping with reference to its input, instead of learning a direct unreferenced mapping, as shown in Fig. 2, middle. The layers of a residual block are formulated as (CONV - BN - ReLU - CONV - BN - ReLU). Assuming that $H(x)$ is a function to be approximated. The residual block learns the residual function $F(x) := H(x) - x$ instead. Thus, the original function to be approximated becomes $F(x) + x$ [10]. We stack ResNet blocks to form a visual feature extraction module of depth 17 CONV layers. This ResNet-17 module is described as {RGB input - (CONV - BN - ReLU - max POOL) - 8 residual blocks - average POOL}. At time step t, it maps an input frame x_t of $224 \times 224 \times 3$ pixels into a fixed length feature vector $f_t \in R^{d=512}$. The resulting feature vectors of an input sequence are passed to the sequence modeling LSTM module.

Due to the scarcity of fall detection datasets [7], and the need for large amounts of data to train deep ConvNet models, we initialise the core ConvNet model from the ResNet-18 model pre-trained by Abobakr et al. [27], [28]. This model was pre-trained on 280K colorised depth images for estimating joint angles of articulated human body postures in workplace environments. This approach provides a strong initialisation that reduces the effect of overfitting due to the small number of videos in fall detection datasets.

C. LSTM Modeling Temporal Dynamics

Recurrent neural networks (RNNs) are powerful at modeling complex temporal dynamics in input sequences [29]. RNN models maintain a hidden state that stores information about previous inputs. At t time step, the RNN maps an input vector x_t to an output vector z_t and an updated hidden state h_t based on the previous hidden state vector h_{t-1} and the current input x_t. Traditional RNN models are challenged to learn long-term temporal dependencies due to vanishing and exploding gradients problem [30].

The Long Short-Term Memory (LSTM) [30] network overcomes these challenges and features easier approach to learning long-term temporal dynamics. In addition to the hidden state, it incorporates a memory unit or cell state that is continuously modified using non-linear gating functions, which are learned. These gating functions manipulates the memory unit through forget and update operations to allow storing only relevant information. The recurrent LSTMs have demonstrated superior capabilities to model long-term dependencies and learn to recognise and synthesise sequences in several application domains such as speech recognition [31], pedestrian behaviour modeling [32], [33] and medical imaging [34].

The sequence modeling module has two stacked LSTM units with hidden states of size 100. These LSTMs learn to map a sequence of feature vectors F $= \{(f_t, y)\}_{t=1}^{T}$, where F $\in R^{T\times512}$, produced using the ResNet visual feature extraction module, to a single output vector Z_T. We use the UR Fall Detection Dataset (URFD) [21] for training and evaluating the proposed method. Qualitative analysis on this dataset suggested that the fall action takes approximately 80 frames. Therefore, we train the ConvLSTM network on video clips of length $T = 80$ depth frames.

The final output of the proposed ConvLSTM model is a probability distribution $P(\hat{y}_T)$ over C classes $\in$ {fall, non-fall}. We pass the last output Z_T to a fully connected layer that performs the logistic regression $\hat{y}_T = W_z Z_T + b_z$, where $W_z \in \mathbb{R}^{|C|\times d_z}$ and $b_z \in \mathbb{R}^{|C|}$ are learned parameters. The softmax function is finally used to normalise the produced logits and compute the posterior probability distribution over fall and non-fall classes as:

$$P(\hat{y}_T = c|Z_T; W_z) = \frac{\exp(\hat{y}_T, c)}{\sum_{k\in C}\exp(\hat{y}_T, k)} \tag{1}$$

where W_z represents model parameters, and the denominator is used for normalisation so that the class scores sum to one. The key aspect of the proposed ConvLSTM model is that all its stacked components: ResNet, LSTM and logistic regression, are trained jointly from an end-to-end using generic optimisation algorithm operating on backpropagated gradients. The training objective function is minimising the binary cross entropy loss over a training mini-batch of sequences:

$$E_{train} = -\frac{1}{N}\sum_{i=1}^{N}\left[t_i \log \hat{y}_T + (1-t_i)\log(1-\hat{y}_T)\right] \quad (2)$$

where N is the mini-batch size in sequences, $\hat{y}_T$ is the predicted posterior probability distribution for the sequence x_i and t_i is the target probability distribution.

IV. Experiments and Results

We have trained an end-to-end ConvLSTM model for fall activity recognition. This model is composed of a ResNet convolutional network for visual features extraction, a stack of two LSTM units for sequence modeling and a logistic regression layer for recognition. The input to our model is a sequence of depth frames and the output is probability distribution over two classes: fall or non-fall. This section evaluates the performance of the proposed fall detection method. First, training and testing datasets are described. Second, we detail the used training parameters. Third, evaluation criteria that are commonly used to assess the performance of fall detection systems are described. Finally, we compare the proposed method with other approaches in the literature.

A. Dataset

We use the UR Fall Detection Dataset (URFD) [21] for training and evaluating the performance of the proposed method. The latest comprehensive comparison of RGB-D datasets lists only URFD and TST datasets under the fall detection category [35]. However, the TST dataset is not currently available for download. So, the experiments have been conducted on the URFD dataset only. It has a total of 70 activities distributed as: 30 falls from standing and sitting on a chair, 30 ADL activities such as walking, sitting down, squatting and picking-up an object and 10 sequences with fall-like activities such as lying on wooden sofa and lying on the floor. These activities were performed by five subjects of different anthropometric measures. We split each of these activity classes into 80% for training and 20% for validation. This setting produces 55 video sequences for training and 15 for validation. Due to this limited number of activities, we performed data augmentation by segmenting the activities using a sliding window of T depth frames and a stride of one frame.

Qualitative analysis has been performed to choose the sequence length T. We concluded that the fall motion takes about 80 frames including an initial period, rapid movement and resting on the ground. Therefore, dataset activities are segmented using a sliding window of 80 frames with stride of one to include variability to the start and ending periods. We randomly selected 1750 sequences for training and 725 sequences for validation. Training and validation sequences are balanced among dataset classes: fall, ADL and fall-like ADL activities.

TABLE I
Evaluation and comparison of SVM fall detection on the URFD dataset. Higher is better.

Method	Accuracy	Precision	Sensitivity	Specificity
SVM + Acc. [21]	0.94	0.88	1.00	0.90
RDF + SVM [7]	0.96	0.91	1.00	0.93
Proposed	**0.98**	**0.97**	**1.00**	**0.97**

The proposed system achieves better results than the accelerometer-based system of Kwolek et al. [21] and the RDF+SVM method of Abobakr et al. [7]. The precision is the ratio between true positives and the sum of both true and false positives.

B. Training Parameters

The components of the proposed ConvLSTM model are trained jointly and end-to-end. We used the stochastic gradient descent (SGD) optimisation algorithm with an initial learning rate of 0.0005, decaying by a factor of 10 every 30 epochs, mini-batch size of 160 frames representing 2 sequences, weight decay of 0.0001 and momentum of 0.9. This model takes approximately 12 hours of training on a NVIDIA Titan X GPU for 100 epochs.

C. Evaluation Criteria

Fall detection sensors performance assessment criteria defined in [36] are used. Noury el al. [36] performed analysis on series of tests and concluded with proposing

$$Sensitivity = \frac{TP}{TP + FN} \quad (3)$$

measuring the capacity to recognise the considered occurrence properly, for instance, a fall motion or a lying posture, and

$$Specificity = \frac{TN}{TN + FP} \quad (4)$$

determining the capacity to correctly detect non-falls as two evaluation metrics, where

- True positive (TP): correctly identifies the occurrence of a lying posture or a fall motion,
- False positive (FP): false alarm indicates that a posture is misclassified as lying or and ADL activity is identified as a fall,
- True Negative (TN): no fall alarm on ADL events and standing or sitting postures,
- False Negative (FN): incorrect identification of a fall.

D. Fall Detection Results

The proposed method achieves an average activity recognition accuracy of 98% on the validation set. To compare with other approaches in the literature, we follow the same protocol used in [21] and [7]. The reported results are on the whole dataset of 70 sequences, on a misclassification event, the rest of the sequence is ignored. Kwolek et al. [21] recorded the dataset and kindly shared it for evaluation and comparison of fall detection methods. Their approach indicates an eventual fall through the thresholding of accelerometric signal and uses a lying posture classifier to authenticate the occurrence of the fall. Abobakr et al. [7] used pixel-wise RDF model to recognise the articulated posture either lying, sitting or standing. The change in lying posture confidence over time is assessed using a SVM classifier to detect falls. Table I compares the performance of the proposed ConvLSTM method with both methods. The precision score reported in this table represents the ratio between true positives and the sum of both true and false positives. The proposed system achieves better results than the accelerometer-based system in [21] and the RDF+SVM method of [7] in accuracy, precision and specificity. Only one false alarm has been detected, as a fall-like ADL activity has been misclassified as a fall.

E. Response Rate

The proposed ConvLSTM model is capable of processing 15 sequences, each of 80 frames, per second on a NVIDIA TITAN-X GPU. This means that the proposed method can provide real-time fall detection performance.

V. Conclusion

This paper proposed a vision based fall detection method. A deep ConvLSTM model composed of a hierarchical visual feature extractor ResNet convolutional model, a recurrent LSTM module for sequence modeling and a logistic regression module for fall action recognition. The overall model is trained end-to-end using backpropagated gradients. The proposed method does not require skeleton tracking or person detection and segmentation. We evaluated the proposed method on the public URFD dataset and reported full sensitivity in detecting fall events and only one false alarm. This results features the proposed method as a fast and reliable solution for hospitals, homes, bathrooms and laborious workplaces.

Acknowledgement

This research was fully supported by the Institute for Intelligent Systems Research and Innovation (IISRI) at Deakin University, Australia.

References

[1] Who, "WHO Global Report on Falls Prevention in Older Age." *Community Health*, p. 53, 2007.

[2] S. Deandrea, E. Lucenteforte, F. Bravi, R. Foschi, C. La Vecchia, and E. Negri, "Risk factors for falls in community-dwelling older people:" a systematic review and meta-analysis"," *Epidemiology*, pp. 658–668, 2010.

[3] M. E. Tinetti, W. L. Liu, and E. B. Claus, "Predictors and prognosis of inability to get up after falls among elderly persons." *JAMA : the journal of the American Medical Association*, vol. 269, no. 1, pp. 65–70, 1993.

[4] R. Igual, C. Medrano, and I. Plaza, "Challenges, issues and trends in fall detection systems," *Biomedical engineering online*, vol. 12, no. 1, p. 66, 2013.

[5] G. Wu and S. Xue, "Portable preimpact fall detector with inertial sensors," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 16, no. 2, pp. 178–183, 2008.

[6] H. Kerdegari, K. Samsudin, A. R. Ramli, and S. Mokaram, "Evaluation of fall detection classification approaches," in *Intelligent and Advanced Systems (ICIAS), 2012 4th International Conference on*, vol. 1. IEEE, 2012, pp. 131–136.

[7] A. Abobakr, M. Hossny, and S. Nahavandi, "A skeleton-free fall detection system from depth images using random decision forest," *IEEE Systems Journal*, vol. PP, no. 99, pp. 1–12, 2018.

[8] G. Demiris, M. J. Rantz, M. A. Aud, K. D. Marek, H. W. Tyrer, M. Skubic, and A. A. Hussam, "Older adults' attitudes towards and perceptions of smart home technologies: a pilot study," *Medical informatics and the Internet in medicine*, vol. 29, no. 2, pp. 87–94, 2004.

[9] E. E. Stone and M. Skubic, "Fall detection in homes of older adults using the microsoft kinect." *IEEE journal of biomedical and health informatics*, vol. 19, no. 1, pp. 290–301, 2015.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[11] K. Buys, C. Cagniart, A. Baksheev, T. De Laet, J. De Schutter, and C. Pantofaru, "An adaptable system for rgb-d based human body detection and pose estimation," *Journal of Visual Communication and Image Representation*, pp. 39–52, 2014.

[12] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman *et al.*, "Efficient human pose estimation from single depth images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 12, pp. 2821–2840, 2013.

[13] H. Haggag, M. Hossny, S. Haggag, J. Xiao, S. Nahavandi, and D. Creighton, "LGT/VOT tracking performance evaluation of depth images," *System of Systems Engineering (SOSE), 9th IEEE International Conference on*, pp. 284–288, 2014.

[14] H. Haggag, A. Abobakr, M. Hossny, and S. Nahavandi, "Semantic body parts segmentation for quadrupedal animals," *IEEE Conference on Systems, Man, and Cybernetics (SMC)*, 2016.

[15] H. Haggag, M. Hossny, S. Nahavandi, and O. Haggag, "An adaptable system for rgb-d based human body detection and pose estimation: Incorporating attached props," *IEEE Conference on Systems, Man, and Cybernetics (SMC)*, 2016.

[16] D. Webster and O. Celik, "Systematic review of Kinect applications in elderly care and stroke rehabilitation," *Journal of NeuroEngineering and Rehabilitation*, vol. 11, no. 1, p. 108, 2014.

[17] H. Haggag, M. Hossny, S. Haggag, S. Nahavandi, and D. Creighton, "Safety applications using kinect technology," in *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2164–2169.

[18] P. Fankhauser, M. Bloesch, D. Rodriguez, R. Kaestner, M. Hutter, and R. Siegwart, "Kinect v2 for mobile robot navigation: Evaluation and modeling," in *Advanced Robotics (ICAR), 2015 International Conference on*. IEEE, 2015, pp. 388–394.

[19] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "3D head tracking for fall detection using a single calibrated camera," *Image and Vision Computing*, vol. 31, no. 3, pp. 246–254, 2013.

[20] Z.-P. Bian, J. Hou, L.-P. Chau, and N. Magnenat-Thalmann, "Fall Detection Based on Body Part Tracking Using a Depth Camera." *IEEE journal of biomedical and health informatics*, vol. 2194, no. c, pp. 1–10, 2014.

[21] B. Kwolek and M. Kepski, "Improving fall detection by the use of depth sensor and accelerometer," *Neurocomputing*, vol. 168, pp. 637–645, 2015.

[22] A. Abobakr, M. Hossny, and S. Nahavandi, "Body joints regression using deep convolutional neural networks," in *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on*. IEEE, 2016, pp. 003 281–003 287.

[23] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," *arXiv preprint arXiv:1301.3572*, 2013.

[24] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 345–360.

[25] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 681–687.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[27] A. Abobakr, D. Nahavandi, J. Iskander, M. Hossny, S. Nahavandi, and M. Smets, "A kinect-based workplace postural analysis system using deep residual networks," in *Systems Engineering Symposium (ISSE), 2017 IEEE International*. IEEE, 2017, pp. 1–6.

[28] A. Abobakr, D. Nahavandi, J. Iskander, M. Hossny, S. Nahavandi, and M. Smets, "RGB-D human posture analysis for ergonomic studies using deep convolutional neural network," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2017, pp. 2885–2890.

[29] J. Donahue, L. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description." *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, p. 677, 2017.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014, pp. 1764–1772.

[32] K. Saleh, M. Hossny, and S. Nahavandi, "Intent prediction of vulnerable road users from motion trajectories using stacked lstm network," in *Intelligent Transportation Systems Conference (ITSC), 2017 IEEE International Conference on*. IEEE, 2017, pp. 327–332.

[33] K. Saleh, M. Hossny, A. H. Hossny, and S. Nahavandi, "Cyclist detection in lidar scans using faster r-cnn and synthetic depth images," in *Intelligent Transportation Systems Conference (ITSC), 2017 IEEE International Conference on*. IEEE, 2017.

[34] M. Attia, M. Hossny, S. Nahavandi, and A. Yazdabadi, "Skin melanoma segmentation using recurrent and convolutional neural networks," in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE, 2017, pp. 292–296.

[35] M. Firman, "Rgbd datasets: Past, present and future," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 19–31.

[36] N. Noury, A. Fleury, P. Rumeau, A. Bourke, G. Laighin, V. Rialle, and J. Lundy, "Fall detection - principles and methods," *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pp. 1663–1666, 2007.