Received: December 27, 2019. Revised: January 23, 2020. 250

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A New Approach for Thermal Vision based Fall Detection Using Residual Autoencoder

Faten A. Elshwemy[1]* **Reda Elbasiony[1]** **Mohamed Talaat Saidahmed[1]**

[1]*Computers and Control Department, Faculty of Engineering, Tanta University, Egypt*
* Corresponding author's Email: faten_elshwemy @ f-eng.tanta.edu.eg

Abstract: This paper focuses on falling of the elderly people which is considered as one of the most critical issue that can face them in their life. To deal with such issue, we propose a new approach named a Spatio-temporal Residual AutoEncoder (SRAE) model. This model is an unsupervised fall detector based on utilizing the deep learning technique to detect falls of the elderly people. Our proposed model uses autoencoder based on convolutional neural network, convolutional long short term memory (ConvLSTM) network, and residual connections to extract spatial and temporal features of videos captured from thermal cameras. The reconstruction error of an autoencoder is used to detect falls recorded in such thermal videos. Furthermore, SRAE model is tested on the publicly available thermal dataset where thermal images conserve the privacy of the elderly under observation which is a very important issue. The obtained results show that the our proposed model detects falls with high receiver operating characteristic area under curve (ROC AUC) (97%) ,and precision recall area under curve (PR AUC) (93%) compared to denoising autoencoder (DAE), convolutional autoencoder (CAE), and convolutional long short term memory autoencoder (CLSTMAE) introduced in the literature.

Keywords: Fall detection, Anomaly detection, Deep learning, Residual network, Convolutional, AutoEncoder, Thermal cameras.

1. Introduction

In recent years, the elderly population rapidly increases over the world. Most elderly people suffer from various forms of dementia diseases, so they need special care [1]. The smart homes are one of the internet of things (IoT) applications that have been used to provide health-care monitoring for the elderly people. Falling is a major risk that affects the quality of elderly people life. Remaining on the ground for a long time as a result of a fall can cause serious health problems that can in sometimes lead to death [2]. Different kinds of sensors such as cameras and wearable sensors can be used in the smart homes in order to monitor a person's body pose or motion to detect the elderly people falls without interfering in their life routines [3, 4]. The fall detection technique is an important service for the elderly healthcare to improve the quality of their life and decrease the costs of monitoring these individuals. To detect a fall early, a reliable fall detection system is necessary to reduce falling post-effects. Since fall is an abnormal activity, the fall detection problem is considered as an anomaly detection problem. Recently, automatic fall detection systems have received special attention from the research community. Most existing approaches used different machine learning and deep learning models for training to identify falls. However they are still suffering from occurring false positive. So, finding a robust fall detection system is the main challenge to avoid false positive and obtain high accuracy. Due to the rarity of occurrence of falls, there may be insufficient training data available for building classification models using classical supervised approaches to handle the fall detection problem. In this paper, Spatiotemporal Residual AutoEncoder based residual network (SRAE), is proposed which is an unsupervised fall detection model. Our unsupervised model is trained

on the normal activities of daily living (ADL) only. The SRAE model consists of two networks: encoder and decoder. The encoder input is a window of adjacent frames of a video that is obtained by applying a temporal sliding window to video frames. The SRAE encoder consists of four 2D convolutional layers, each convolutional is followed by the batch normalization (BN) and relu activation function except the first one is followed only by relu activation function. Then, it is followed by one ConvLSTM layer. While the SRAE decoder is built in a symmetrical way with respect to the structure of the encoder. The tanh activation function is used in the last layer of the decoder. To facilitate convergence and enhance the results, the residual connections are added from two encoder layers to their corresponding layers in the decoder. The objective function to train the SRAE model aims to reduce the reconstruction error between input and output. To train and test the proposed model, the infrared thermal dataset that captured from thermal camera is used. Whereas the thermal cameras in smart houses may be preserve the person privacy and capture images during night conditions as well. When the training phase is successfully completed, the proposed model SRAE will learn both the spatial and temporal features.

To summarize, the main contributions of this paper aim to:

- propose a spatiotemporal residual autoEncoder based residual network (SRAE) which is an unsupervised fall detection model,
- test the proposed model on the publicly available thermal dataset where thermal images conserve the privacy of the elderly, and
- compare the proposed model with denoising autoencoder (DAE), convolutional autoencoder (CAE), and convolutional long short term memory autoencoder (CLSTMAE) models.

The remainder of this paper is organized as follows: in section 2 we provide an overview of pervious work in anomaly detection and fall detection. Section 3 presents an overview of the approach model. Section 4 reports the experiments conducted and analyzes the results. Section 5 overviews the conclusion and future work of the paper.

2. Related work

In the last decade, several approaches have been proposed for anomaly detection in various applications. This section reviews some related works in the field of anomaly detection and fall detection which are closer in concept to the work presented in this paper. Authors in [5] propose a deep leaning model for the animal health monitoring called GRU-AE to detect a respiratory disease in the growing pigs. In this model, the gated recurrent units (GRUs) are used to build the autoencoder. The GRU-AE model is trained on data collected from different sensors. The anomaly detector based on threshold is used after the reconstruction errors of all frames are computed to identify which frames are considered as anomalies. In [6], authors propose an unsupervised learning model based on deep learning called Appearance and Motion DeepNet (AMDN) to detect anomalous events in videos. This model consists of two stages. The stacked denoising autoencoders (DAE) are used in the first stage to extract appearance and motion features. While in the second stage, the multiple one-class support vector machine (SVM) models are used to compute a set of anomaly scores using the appearance and motion features that are extracted in the first stage. In [7], convolutional autoencoder (CAE) is used for anomaly detection in videos. The reconstruction error of CAE is used as an anomaly score in each frame of videos. The ConvLSTM-AE in [8] is proposed for detecting anomalies in crowded scenes videos. It consists of two parts spatial encoder and decoder and temporal encoder and decoder. The spatial encoder is built using convolutional layers, and the temporal encoder is built using ConvLSTM layers. Authors in [9] use this ConvLSTM-AE model to detect falls of elderly people using videos captured from thermal cameras. They also propose a new method to score anomalies by combining a reconstruction error of a frame across different sequences of a video. In [10], authors develop the deep learning model to solve the problem of fall detection on video kinematic data. Their model is based on using convolutional neural network and long short term memory (LSTM). It comprises of eight 3D convolutional layers and five pooling layers. Each 3D convolutional layer is followed by 3D pooling layer except the third connected directly to fourth convolutional layer and the fifth connected directly to sixth convolutional layer. In the last layer, the long short term memory (LSTM) layer is used. CNN-3B3Conv is proposed in [11] for human fall detection based on convolutional neural network. This model consists of three sequential blocks. The first & second blocks consist of three convolutional layers followed by pooling layer. They are different only with the kernel size. The last block consists of three fully connected layers with 64, 32, and 2 neurons respectively. In [12], authors propose convolutional neural networks for the fall detection problem. They used the modified version of VGG-

16 CNN model. They use transfer learning technique to classify fall from non-fall activities. The stack of optical flow images is used as the input to the CNN model. A fall detection system based on deep learning and multi-sensor fusion was proposed in [13]. The continuous wave radar and an optical camera are used to obtain the signals of human motions, and capture the images sequence of the human. This system extracts the TF features from the radar signals using the STFT. Three convolutional neural networks are used, in which two of these networks are trained to classify the TF images, and one of them is trained to predict the bounding box variation of the image sequence. Finally, the fall is detected using jointly decision of the result of the three CNNs. In [14], the novel generative model based abnormal event detection method (STAN) is proposed. It consists of the spatio-temporal generator and the spatiotemporal discriminator. They were trained in the adversarial way to represent the spatiotemporal features of normal patterns. In [15], FallDroid application is developed which is a fall detection application based on an Android to monitor elderly people and detect falls. This application uses different carrying location such as thigh and around the waist for smart phones. It uses two step algorithm that consists of threshold based method (TBM) and multiple kernels learning support vector machine (MKL-SVM). In the first step, TBM algorithm discards most of the daily living activities (ADL) using threshold value. While, the second step uses the MKL-SVM algorithm to classify falls.

These previous approaches use different sensors and different methods in various anomaly detection applications. In the context of fall detection approaches, different types of sensors are used to monitor the daily living activities (ADL) of elderly people such as wearable sensors, ambient-based sensors, and vision-based sensors [16]. In the case of wearable sensors, elderly people have to wear sensors all the time, which causes the inconvenience of them. As well, these people often forgot to wear such sensors. While the disadvantages of ambient-based sensors are limited accuracy, and causing a high rate of false alarm. For vision based sensors, although cameras can provide all the information about elderly, they breach people privacy. To preserve the privacy, some cameras such as thermal camera can be used. Also, thermal cameras can capture images during the night and in dark places. So, in this paper, we focus on detecting falls of elderly people in videos captured from thermal cameras.

3. Methodology

In this section, we first briefly review the technologies that we use in our model such as autoencoder, convolutional long short term memory (ConvLSTM), and deep residual networks. Then, we introduce the design of the proposed SRAE model.

3.1 Autoencoder

Autoencoder (AE) is an unsupervised learning algorithm which consists of two network parts: encoder and decoder [17]. The structure of the encoder and decoder are a symmetric. An encoder network is used to learn efficient features of input data and compress it into compression form. While a decoder network is used to reconstruct the original input from this compression form. The objective function of the autoencoder aims to minimize the reconstruction error between the input of the encoder and the output of the decoder using Eq. (1). The backpropagation is used in minimizing the reconstruction error of the autoencoder.

$$ObjFunc = min(\,F(x\,,y)\,) \qquad (1)$$

where x is an input, y is an output, and F is a mean squared error (MSE) function.

3.2 Convolution LSTM network

Long Short Term Memory (LSTM) is a variant of the recurrent neural network that has feedback connection and is capable learning long-term dependencies [18]. It consists of three gates: input gate, forget gate and output gate. These gates are responsible for regulating the flow of information. LSTM contains a cell state that is used to store information to prevent gradient from vanishing. The formulation of the LSTM unit can be summarized with the equations from (2) to (5) [18].

$$g_t^j = \sigma\big(W_{xj}x_t + U_{hj}h_{t-1} + b_j\big) \qquad (2)$$

$$\widehat{c_t} = tanh(W_{xc}x_t + U_{hc}h_{t-1} + b_c) \qquad (3)$$

$$c_{\ t} = c_{t-1} \circ g_t^f + \widehat{c_t} \circ g_t^i \qquad (4)$$

$$h_t = g_t^z \circ \tanh(c_t) \qquad (5)$$

where g^jstands for input gate activation vector ($i \in \mathbb{R}^h$), forget gate activation vector ($f \in \mathbb{R}^h$), and output gate activation vector ($z \in \mathbb{R}^h$), $x_t \in \mathbb{R}^d$denotes to input vector to the LSTM unit, $h \in \mathbb{R}^h$

denotes to hidden state vector, $c \in \mathbb{R}^h$ denotes to cell state vector, $\hat{c}_t$ denotes to update of the cell gate, $b \in \mathbb{R}^h$ is the bias, tanh is the hyperbolic tangent function, and σ is the sigmoid function. $Ws \in \mathbb{R}^{h \times d}$, and $Us \in \mathbb{R}^{h \times h}$ are weight matrices connecting various components. d and h refer to the number of input features and number of hidden units, respectively. The symbol ∘ denotes element-wise multiplication operation. The advantage of LSTM is its ability in modeling temporal dependencies of sequences. Although, LSTM suffers from ignoring the spatial information for multi-dimensional sequence data. So, Convolutional Long Short-term Memory (ConvLSTM) model was introduced by Shi et al. in [19] which is a variant of the LSTM architecture to overcome this drawback. The equations of ConvLSTM unit are similar to LSTM unit; the input is fed in as images. The matrix operations in ConvLSTM model are replaced with convolution operation (the symbol ∗ denotes a convolution operation). ConvLSTM units require fewer weights by using convolution for both connections between input and hidden layers and between hidden and hidden layers. This allows ConvLSTM to propagate well spatial features temporally through each ConvLSTM state with images than the LSTM unit. The formulation of the ConvLSTM unit can be summarized with the equations from (6) to (9) [19].

$$g_t^j = \sigma\left(W_{xj} * x_t + U_{hj} * h_{t-1} + b_j\right) \quad (6)$$

$$\hat{c}_t = tanh(W_{xc} * x_t + U_{hc} * h_{t-1} + b_c) \quad (7)$$

$$c_t = c_{t-1} \circ g_t^f + \hat{c}_t \circ g_t^i \quad (8)$$

$$h_t = g_t^o * tanh(c_t) \quad (9)$$

3.3 Deep residual network

The most problem with training deep learning is the degradation problem. This problem happens because of the training accuracy is saturated and then decreased quickly with increasing of the network depth. To solve this problem, the residual network is proposed. ResNet is the convolutional neural network (CNN) model [20] which makes training of the deep neural network efficiently by solving the degradation problem. This is done by using skip connection or shortcut over the network layers. In the residual blocks of a network, each output layer is fed into the input of the next layer and directly into the input of a layer about 2–3 jumps as shown in Fig. 1. A residual learning block can be defined as in Eq. (10) [20].

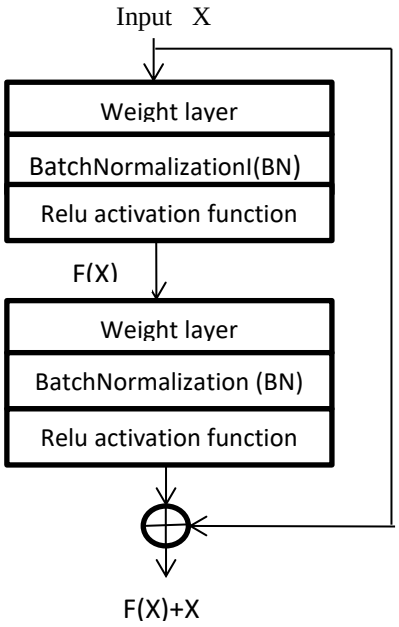

Fig.1. Residual learning block

$$Y = F(X) + X \quad (10)$$

where X is the input tensor, F can be several convolutional layers, and Y is the output tensor of the residual block.

3.4 Spatio-temporal residual autoencoder

In this section, we present our proposed model used for fall detection problem. Our proposed model SRAE is an unsupervised fall detection model using the convolutional autoencoder based on the residual connections, Convolution2D, and ConvLSTM, to detect falls as anomalies. The structure of the SRAE model is shown in Fig. 2. As can be seen, it consists of the encoder and decoder networks. The encoder consists of four 2D convolutional layers. The first convolutional layer is followed by ReLU activation function. The other convolutional layers are followed by batch normalization layer and rectified linear unit (ReLU) activation function. In the encoder part, the down sampling is performed by convolutional layers using stride. While in the decoder part, the up sampling, which is the inverse operation of the down sampling, is performed by transposed convolutional layers to increase the resolution of the output. The kernel size, the number

Table 1. Configuration of the proposed SRAE model for encoding and decoding phases

	Layer	Kernal	Stride	Filters
Encoder	Conv2D	7x7	2 x 2	128
	Conv2D	5x5	2 x 2	64
	Conv2D	5x5	2 x 2	64
	Conv2D	5x5	2 x 2	64
	ConvLSTM2D	3x3	1 x 1	32
Decoder	ConvLSTM2D	3x3	1 x 1	32
	Conv2DTranspose	5x5	2 x 2	64
	Conv2DTranspose	5x5	2 x 2	64
	Conv2DTranspose	5x5	2 x 2	64
	Conv2DTranspose	7x7	2 x 2	1

of filters, and the stride value applied for each layer of the encoder and decoder are shown in Table 1. The batch normalization is used to perform faster training and reduce the over fitting. We use the rectified linear unit activation function in order to improve the gradient propagation through the training phase to avoid the problems of exploding and vanishing gradient. Then is followed by ConvLSTM layer because it captures spatial and temporal correlations better. Finally, the decoder has a symmetric structure of the encoder with residual connections between encoder and decoder layers which generates the output images with the same input size. The residual connection is used to facilitate convergence. As the difficulty to collect fall datasets because collecting falls data can put life of people in danger, we train the SRAE model in unsupervised way in which, it is trained using only the normal activities of daily living (ADL). While, in the testing phase, SRAE model is tested on both the normal and abnormal activities. Windows of contiguous video frames are fed to SRAE model by applying a temporal sliding window method given in [9] on all video frames. The number of windows is calculated as in Eq. (11).

$$W = \frac{no\ of\ frames - L}{S} + 1 \tag{11}$$

where W is the total number of windows, L is the length of the window and S is the stride. In our implementation of SRAE, we set L = 8 and S = 1. The input window enters as input to the SRAE model that encodes it into spatiotemporal features, and then the decoder reconstructs it from the encoded representation. We use mean squared error loss function (MSE) between the input and the reconstructed output for training SRAE model as in Eq. (12):

$$RE_i = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2 \tag{12}$$

where RE is a reconstruction error, n is the number of training samples, X is an input window and Y is an output window.

To detect falls, we compute the reconstruction error for every frame across different windows. Then we compute the average of all reconstruction errors of the frame across different windows because the same frame may appear in multiple windows as shown in Eq. (13). This average is used as anomaly score. When the average of reconstruction error per frame is low, this indicates the normal activity because the reconstruction error of a normal activity frame should not vary a lot with its position across different windows. When the average of reconstruction error per frame is high, this indicates the occurrence of a fall.

$$AvgRE_j = \frac{1}{w}\sum_{i=1}^{w} RE_{ij} \tag{13}$$

where $AvgRE_j$ is the average of reconstruction error for a frame j, w is number of windows where the frame j is appeared in, and RE_{ij} is a reconstruction error of a frame j in window i.

4. Experimental methodology and results

4.1 Dataset

The proposed model is tested on the thermal fall dataset [2]. It consists of normal activities of daily living and fall videos captured by a FLIR ONE thermal camera. This camera is mounted on an Android phone in a room where setting with a single view with a spatial resolution of 640 × 480. The total number of videos is 44 which 9 videos contain ADL only and 35 videos contain falls with normal ADL. This dataset has different positions for falls occurrence such as falling from standing, falling

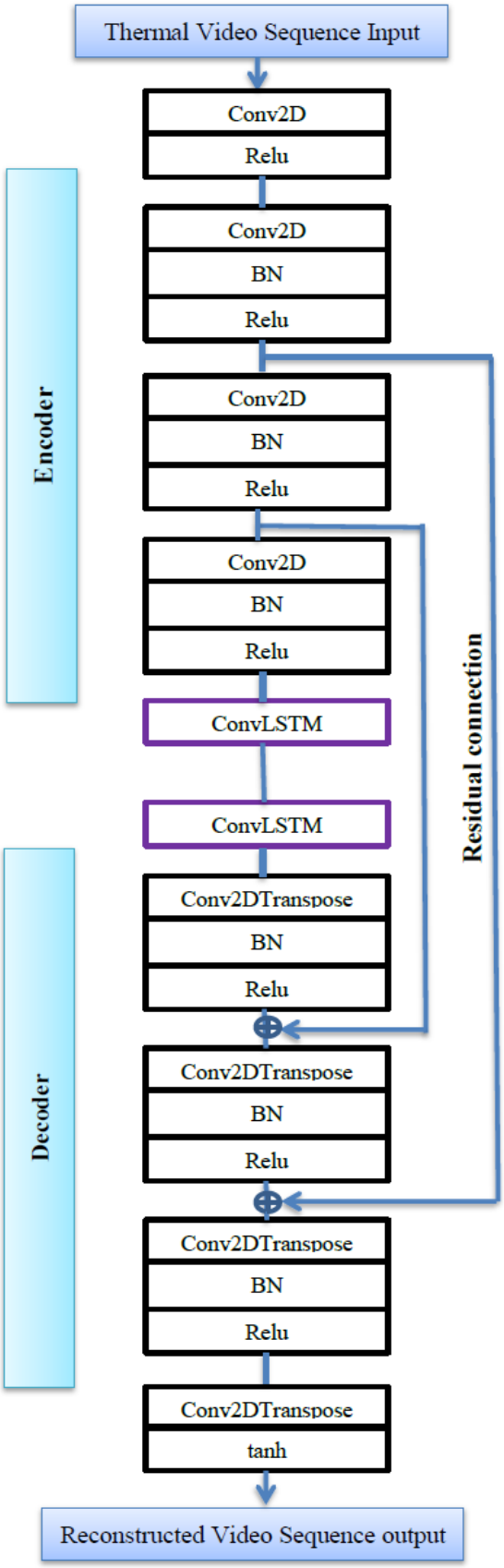

Figure.2 Proposed RSAE model

from chair, and falling from sitting. Samples of frames containing thermal normal activities of daily living (ADL) and fall are shown in Fig. 3. The sliding window is applied on all frames of videos as explained in the previous section to create windows of contiguous video frames to give as input to the SRAE model for training and testing. In the thermal fall dataset, the total number of frames is 22, 116 daily living activities frames from 9 videos. After windowing the ADL videos individually using Eq. (11), 22, 053 windows of adjacent frames are obtained, which are used for training the proposed model.

(a)

(b)

Figure.3 Thermal dataset – ADL frames: (a) normal ADL and (b) fall frames

4.2 Experimental setup and results

The experiments aim to evaluate the performance of the proposed SRAE model compared to other deep learning works. We ran our experiments on a processor core i7 (4GHz, 8 M cache) with 16 GB RAM and GPU TITAN X 12 GB running Windows 10. We implemented all models and training procedures in Keras with Tensorflow backend. The frames of the thermal dataset are extracted from video files. We compared the proposed model with denosing autoencoder (DAE), convolutional autoencoder (CAE), and convolutional long short term memory autoencoder (CLSTMAE) because they are applied in [9] on the same thermal dataset that we are used. The configurations of the encoding and decoding phases for the DAE, CAE, and CLSTMAE models are illustrated in Table 2 [9]. We use the same results for the compared models as in [9]. The CAE and DAE models are trained for 200 epochs. While CLSTMAE model is trained for 50 epochs [9]. The proposed model is trained for 30 epochs. The Adadelta optimizer was used in training all models. The training batch size is set to 32 for all models except for CLSTMAE model is set to 16. Every batch consists of windows of 8 frames. All thermal frames are resized to 64 × 64. The input to the network is a single frame for all compared models, while the proposed and CLSTMAE models, the input to the network is a window of adjacent frames. When our proposed model is trained, we do not use any data augmentation. We trained our model on only normal ADL videos. While, in the testing phase, our model is tested on fall videos that contains of both normal and fall frames. A reconstruction error per frame, for all models except the proposed and CLSTMAE, is calculated that can be used as anomaly score to identify a fall frame as an anomaly. For the proposed and CLSTMAE, A reconstruction error per frame across different windows for a given video is computed, which can be used as anomaly score to identify a fall frame as an anomaly. We train all models for 200 epochs to show the convergence rate for each model. The convergence curves for all models are shown in Fig. 4. This figure shows that the proposed SRAE model has fastest convergence rate than other compared models.

The comparisons are made in terms of Receiver Operating Characteristic area under curve (ROC AUC) and Precision Recall area under curve (PR AUC). A ROC curve is plotting true positive rate (TPR) as shown in Eq. (14), which is referred to as sensitivity, against false positive rate (FPR) as shown in Eq. (15), which is referred to specificity. The ROC AUC is the area under the ROC curve. The higher ROC AUC is, the better the model is. While, a PR curve is plotting precision rate as shown in Eq. (16) against recall rate as shown in Eq. Eq. (17). Recall is the same as sensitivity. The PR AUC is the area under the PR curve. The higher PR AUC is, the better the model is.

$$TPR = \frac{True\ Positives(TP)}{(True\ Positives(TP)+False\ Negatives(FN))} \quad (14)$$

$$FPR = \frac{False\ Positives(FP)}{(False\ Positives(FP)+True\ Negatives(TN))} \quad (15)$$

$$Precision = \frac{True\ Positives(TP)}{(True\ Positives(TP)+False\ Positives(FP))} \quad (16)$$

$$Recall = \frac{True\ Positives(TP)}{(True\ Positives(TP)+False\ Negatives(FN))} \quad (17)$$

The results for all models: DAE, CAE, CLSTMAE, and the proposed SRAE are shown in Table 3. The AUC results are the average of AUC across all test videos, with standard deviation in brackets. We observe that CAE model performs better than DAE model. This is because; DAE fails to extract the 2D spatial features of images by flatting the image input into 1D vector of pixels. However, CAE can extract the 2D spatial features of images. We also observe that CLSTMAE, and the proposed SRAE models perform better than CAE model. This is because; CLSTMAE and SRAE models extract both the spatial and temporal features in videos which are important in fall detection. It should be noted that SRAE model performs better than CLSTMAE model due to use residual connections that improve the performance as mentioned in previous sections. Interesting to see also that, our model SRAE holds the highest AUC values. Its improvements of ROC AUC and PR AUC are 14 % and 24% than CLSTMAE model, 22 %, and 74% than CAE model, and 33%, and 76% than DAE model, respectively.

Table 2. Configurations of encoding and decoding phases for DAE, CAE, and CLSTMAE models

	DAE	CAE	CLSTMAE
Encoder	Dense(4096) Dense(1500) Dense(1000) Dense(500)	Conv 2D kernal 3 x 3, 16 filters Max_pooling 2D (2,2) Conv2D kernal 3 x 3, 8 filters Max_pooling2D (2, 2) Conv2D kernal 3 x 3, 8 filters Max_pooling2D (2,2)	Conv2D kernal 11 x 11, 128 filters , stride=4 Conv2D kernal 5 x 5, 64 filters , stride =2 ConvLSTM2D kernal 3 x 3, 64 filters ConvLSTM2D kernal 3 x 3, 32 filters ConvLSTM2D kernal 3 x 3, 64 filters
Decoder	Dense(500) Dense(1000) Dense(1500) Dense(4096)	Conv2dTranspose kernal 3 x 3, 8 filters Up_sampling2D (2, 2) Conv2D Transpose kernal 3 x 3, 8 filters Up_sampling2D (2, 2) Conv2D Transpose kernal 3 x 3, 16 filters Up_sampling2D (2, 2) Conv2D Transpose kernal 3 x 3, 1 filter	Conv2DTranspose kernal 5 x 5, 128 filters , stride=2 Conv2DTranspose kernal 11 x 11, 1 filter , stride =4

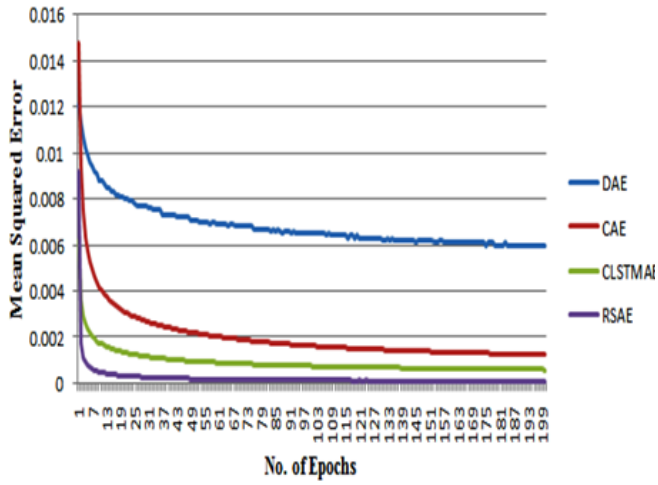

Figure.4 Convergence curves for DAE, CAE, CLSTMAE, and proposed RSAE models with 200 epochs

Table 3. AUC values for different models for Thermal dataset

Model	ROC AUC	PRAUC
DAE	0.64 (0.15)	0.17 (0.15)
CAE	0.75 (0.16)	0.19 (0.18)
CLSTMAE	0.83 (0.11)	0.69 (0.15)
SRAE	**0.97 (0.08)**	**0.93 (0.13)**

5. Conclusion and future work

In this work, a spatiotemporal residual autoencoder (SRAE) is proposed to solve the problem of fall detection as anomaly detection of the elderly people. Our proposed model is trained in an unsupervised manner. It is trained only on normal activities of daily living (ADL). The proposed model extracts the spatial and temporal features of thermal videos using convolutional layers and convolutional long short term memory (ConvLSTM) layers. To improve the training accuracy, we used residual connections between encoder and decoder layers. In the test phase, the proposed model is tested on both normal ADL and fall actions. We used the infrared thermal fall dataset to test the proposed model. To estimate the performance of the proposed model, we compared it with some previous deep learning approaches. The comparison results show that our proposed model has high performance and accuracy to detect falls than DAE, CAE, and CLSTMAE models. The proposed SRAE model holds the highest ROC AUC (97 %) and PR AUC (93 %) values than other compared models: DAE (ROC AUC 64 % and PR AUC 17 %), CAE (ROC AUC 75 % and PR AUC 19 %), and CLSTMAE (ROC AUC 83% and PR AUC 69%). In this paper, the main application of the proposed SRAE model is fall detection. For future work, we plan to apply the proposed model to different anomaly detection applications. Also, we plan to apply the proposed model on different types of datasets such as depth camera datasets.

References

[1] R. Igual, C. Medrano, and I. Plaza, “Challenges, Issues and Trends in Fall Detection Systems”, *Biomedical Engineering Online*, Vol.12, No.66, 2013.

[2] S. Vadivelu, S. Ganesan, O. R. Murthy, and A. Dhall, “Thermal Imaging based Elderly Fall Detection”, In: *Proc. of ACCV Workshop, Springer*, pp. 541–553, 2016.

[3] D. Yacchirema, J. Suárez de Puga, C. Palau, and M. Esteve, “Fall Detection System for Elderly People using IoT and Big Data”, In: *Proc. of the 9th International Conference on Ambient Systems, Networks and Technologies*, pp.603-610, 2018.

[4] C. Ma, A. Shimada, H. Uchiyama, H. Nagahara, and R. Taniguchi, "Fall Detection using Optical Level Anonymous Image Sensing System", *International Journal of Optics and Laser Technology, Elsevier*, Vol.110, pp.44-61, 2019.

[5] J. Cowton, I. Kyriazakis, T. Plötz, and J. Bacardit, "A Combined Deep Learning GRU-Autoencoder for the Early Detection of Respiratory Disease in Pigs Using Multiple Environmental Sensors", *Sensors*, Vol. 18, No. 2, 2018.

[6] M. Ribeiro, A. E. Lazzaretti, and H. S. Lopes, "A Study of Deep Convolutional Auto-encoders for Anomaly Detection in Videos", *Pattern Recognition Letters*, Vol.105, pp.13-22, 2018.

[7] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting Anomalous Events in Videos by Learning Deep Representations of Appearance and Motion", *Computer Vision and Image Understanding*, Vol. 156, pp. 117–127, 2017.

[8] Y. Chong and Y. Tay, "Abnormal Event Detection in Videos Using Spatiotemporal Autoencoder", In: Cong F., Leung A., Wei Q. (eds) Advances in Neural Networks, *Lecture Notes in Computer Science*, *Springer, Cham*, Vol.10262, pp.189-196, 2017.

[9] J. Nogas, S. Khan, and A. Mihailidis, "Fall Detection from Thermal Camera Using Convolutional LSTM Autoencoder", In: *Proc. of the 2nd workshop on Aging, Rehabilitation and Independent Assisted Living, IJCAI Workshop,* Sweden, 2018.

[10] N. Lu, Y. Wu, L. Feng, and J. Song, "Deep Learning for Fall Detection: Three-Dimensional CNN Combined With LSTM on Video Kinematic Data", *IEEE Journal of Biomedical and Health Informatics*, Vol. 23, No. 1, pp. 314-323, 2019.

[11] G. Santos, P. Endo , K. Monteiro, E. Rocha, I. Silva, and T. Lynn, "Accelerometer-Based Human Fall Detection using Convolutional Neural Networks", *Sensors*, Vol.19, No.7, 2019.

[12] A. Nu´nez-Marcos, G. Azkune, and I. Arganda-Carreras, "Vision-based Fall Detection with Convolutional Neural Networks", *Wireless Communications and Mobile Computing*, 2017.

[13] X. Zhou, Q. Li-Chang, Y. Peng-Jie, D. Ze-Gang, and H. Yu-Qi, "Fall Detection using Convolutional Neural Network With Multi-Sensor Fusion", In: *Proc. of IEEE International Conference on Multimedia & Expo Workshops*, pp.1-5, 2018.

[14] S. Lee, H. Kim, and Y. Ro, "Stan: Spatio-Temporal Adversarial Networks for Abnormal Event Detection", In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.1323–1327, 2018.

[15] A. Shahzad and K. Kim, "FallDroid: An Automated Smart-Phone-Based Fall Detection System using Multiple Kernel Learning", *IEEE Transactions on Industrial Informatics*, Vol.15, No.1, pp.35- 44, 2019.

[16] Y. Birku and H. Agrawal, "Survey on Fall Detection Systems", *International Journal of Pure and Applied Mathematics*, Vol.118, No.18, pp.2537–2543, 2018.

[17] D. Charte, F. Charte, S. García, J. Jesus, and F. Herrera, "A Practical Tutorial on Autoencoders for Nonlinear Feature Fusion: Taxonomy, Models, Software and Guidelines", *Journal of Information Fusion*, Vol.44, pp.78-96, 2018.

[18] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", *Neural Computation*, Vol. 9, No. 9, pp.1735-1780, 1997.

[19] X. Shi, Z. Chen, H. Wang, D. Yueng, W. Wong, and W. Woo, "ConvolutionalLSTM Network: A Machine Learning Approach for Precipitation Nowcasting", In: *Proc. of the 28th International Conference on Neural Information Processing Systems: Curran Associates*, Vol.1, pp.802-810, 2015.

[20] Y. Zhang, W. Chan, and N. Jaitly, "Very Deep Convolutional Networks for End-to-End Speech Recognition", *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.4845–4849, 2017.