Journal of Ambient Intelligence and Humanized Computing (2022) 13:5751–5762
https://doi.org/10.1007/s12652-021-03250-5

ORIGINAL RESEARCH

Vision based human fall detection with Siamese convolutional neural networks

S. Jeba Berlin[1] · Mala John[1]

Received: 16 November 2020 / Accepted: 29 March 2021 / Published online: 12 April 2021

Abstract
Fall detection is drawing serious attention all across the globe, as unattended fall of senior citizens creates long lasting injuries. This necessitates the deployment of automatic fall detection systems to facilitate smart care health environments for the elderly people living in various settings, viz., living independently in their homes, hospitalized or living in care homes. The proposed work employs Siamese network with one shot classification for human fall detection. Unlike the neural network that classifies the video sequences, this network learns to differentiate the video sequences by computing the similarity score. The network contains two identical CNNs, receiving pair of video sequences as the input. The features of these networks are merged at the final layer through the similarity function. Two different architectures viz., one with 2D convolutional filters and the other with depth wise convolutional filters, each operated on two set of features, RGB and optical flow features are developed. Experimental results demonstrate the effectiveness and feasibility of the proposed work compared to state-of-the methods.

Keywords Fall detection · Assisted living · PWCNet · Siamese network · One shot classification

1 Introduction

With the increase in older population, assisted home environments are flourishing worldwide to automatically monitor the human activities so as to help the elderly people living alone and to reduce their health care costs. Typically, fall detection is considered as an emergent technique in elderly monitoring system due to increased risks of injuries and deaths entailed by falls, often for the adults more than 65 years (Deandrea et al. 2010). According to WHO, it is estimated that 646,000 among 37 millions falls are fatal, as recorded in a year. The visual and muscle impairment, frequent loss of balance, dizziness, medication side effects, unconsciousness and slipping (Rubenstein 2006) are the major causative factor of falls.

Moreover, staying over the floor after falls for a prolonged period and the delayed medical assistance may increase the risk of both physical and psychological complications of the elderly people (Ozcan et al. 2017). Therefore, it is necessary to have an automatic fall detection system to rigorously monitor the health status of the hospitalized patients, retirement homes and nursing home residents and, consequently alarm signals (Shieh and Huang 2012; Yacchirema et al. 2019) are sent to the caregivers to provide proper assistance inorder to increase their survival rate.

In recent past, the deployment of human fall system is becoming a hot research problem all across the globe. Existing fall detection systems are categorized into wearable device based systems and context aware systems (Vallabh and Malekian 2018; Jansi et al. 2020). Basically, a wearable device based system (Kerdjidj et al. 2020) utilizes the sensors such as accelerometer and gyroscope for fall detection. This sensing mechanism is cost effective and offers privacy to the users, but the limitation is that the user has to continuously wear the device and it needs frequent charging of batteries.

Compared to wearable sensors, the computer vision based fall detection gives complete information regarding the posture, walking speed, gait pattern, position and location of the user without human intervention. The vision-based system could be based on single/multiple calibrated cameras. Basically, the usage of multiple cameras offers three dimensional data, but still the calibration is complicated, costly and time

✉ S. Jeba Berlin
jebaberlin@gmail.com

[1] Department of Electronics Engineering, Madras Institute of Technology, Anna University, Chennai, India

consuming. However, vision based monitoring is confronted by challenges arising due to occlusion, variation in illumination, dynamic backgrounds and viewpoint variations. Further, the vision based system compromises on the privacy.

Most of the fall detection is carried out using simple shape related features and motion features (Iazzi et al. 2020). The shape information is affected by camera view angle and occlusion. In addition, some of the existing methods use silhouette area (Zerrouki et al. 2018; Nunez-Marcos et al. 2017). The performance of this scheme is affected with variation in size, occlusion and scale variation arising due to the movement of the objects towards or away from the camera. However, fall detection in an outdoor environment has additional complexity associated with locating and detecting the humans. This could be overcome using region proposal network. It (Yao et al. 2020) is basically a convolutional network which is trained end-end manner, exclusively to generate the proposal for the object of interest. It employs pyramidal structure to predict the region proposals of different scales and aspect ratios efficiently.

In recent past, the deep learning neural networks have achieved great success in machine learning and computer vision tasks due to its promising performance (Yao et al. 2020; Khraief et al. 2020). These networks rely on huge data sample to train the network to have good generalization capability. However, Siamese network has a great generalization capacity, though the network is trained with limited examples per class label. One of the fascinating natures of Siamese network is that, only one data sample is sufficient to train the model. However, the training complexity of the network increases as the number of class label increases owing to the nature of pair-wise input data (Zhang et al. 2019a, b). But, the training time is of no concern in the proposed work which is attributed to binary classification problem. A Siamese network does not require any pre-processing procedures and it is proven that the network yields promising performance in the applications such as image retrieval, visual tracking and face recognition (Zhang et al. 2019a, b, 2020).

Siamese network is used to perform distance metric based end-end learning. This network indeed learns significant spatio temporal features that aid in distinguishing a different class labels (Leal-Taixe et al. 2016). The network uses shared weights, thus requiring same number of network parameters, but with twice the computation (Taigman et al. 2014). Siamese network has a great potential on image recognition tasks especially in face recognition, based on judging the image pairs. In addition, Siamese network exploits online process to circumvent time consuming problem in real time tracking. The network has drawn special attention in visual tracking due to its balanced accuracy and speed (Wang et al. 2020). Moreover, this network plays a vital role in offline signature verification on a writer independent context (Ruiz et al. 2020). Motivated by the facts stated above, Siamese network has been employed for 'fall detection' in the proposed work. In this work, Siamese network that operates on RGB and optical flow features with end-end learning mechanism is utilized.

The organization of this paper is as follows. The most recently reported fall detection techniques are summarised in Sect. 2. Section 3 describes the proposed method for human fall detection. The experimental results and discussion are presented in Sect. 4 and finally, the conclusions are provided in Sect. 5.

2 Related work

Some of the works related to vision based fall detection techniques are discussed in detail in this section.

The simple conventional technique in human fall detection is the use of human silhouette information from the given sequences for characterizing human falls. The video sequences could be captured through depth camera (Ma et al. 2014), infra red camera (Mastorakis and Makris 2014) or RGB camera. For the depth based images, the thresholding is done to obtain the silhouette of an image. However, in RGB images, the silhouette of the image is extracted by means of modified running average method (Mirmahboub et al. 2013), codebook based background generation (Yu et al. 2012), frame differencing (Liu et al. 2010) and background subtraction (Abobakr et al. 2018; Fan et al. 2019) techniques.

After the silhouette extraction, the elliptic fitting or rectangle fitting bounding boxes (Soni and Choudhary 2018; Iazzi et al. 2020; Min et al. 2018) are generated. Later, the simple, but effective features such as height-width ratio, height width differences, centre variation rate, effective area ratio and critical time difference are computed. But, in some cases to make the system suitable for real scenarios with reduced computation time, orientation angle and Hu moment invariants (Joshi and Nalbalwar 2017) are considered.

Juang and Chang (2007) projected human silhouettes onto vertical and horizontal axis followed by histogram computation. This projected histogram is further used for the computation of Fourier coefficients, to provide the frequency related information about human falls. In Zerrouki and Houacine (2018), the curvelet transform is applied on silhouette image to provide structural and directional information in frequency domain employing multiple radial directions. The curvature scale space features extracted from the silhouette of each individual frames are considered to be invariant to scale, rotation, translation and action length changes.

Inspite of considering the complete silhouette information, few points in a silhouette are also employed to represent the human shape analysis. Further, fall detection is also made possible through computation of canny edge points (Rougier et al. 2011), followed by shape matching to quantify the shape deformation. In Shieh and Huang (2012), one-pixel wise thinning of edges is done and then, the sequence of characters is generated, according to the direction of the neighbouring pixel to represent the characteristics of human shape on each posture.

Fitting an ellipse alone is insufficient for the effective representation of the human postures. Therefore, the silhouette related features are combined with the global features including motion history image (Rougier et al. 2011), accumulated image map (Bhavya et al. 2016) and integrated normalized motion energy image map (Feng et al. 2014), which are widely used to represent the holistic motion and speed of movement in human falls.

After the initial judgement of human falls based on height-width changes of the contour, the speed and position of the human is seldom determined using optical flow (Iazzi et al. 2020) and the human upper part of the body respectively to enhance the performance of the system. Considering the position of head points, the temporal change of head position, angle of fall and speed of fall of head are also utilized as the feature for classification of falls (Lotfi et al. 2018; Yau et al. 2020). The other effective feature used is 3D head trajectories (Rougier et al. 2013) which are captured using shape and color information of the hierarchical particle filter. Subsequently, it employs velocity characterization to detect falls. This 3d representation is made possible even with the single calibrated camera to make the system pose invariant and cost effective.

Prior to the final level of judgement on fall classification, the first level of segregation between fall like behaviours from daily activities is carried out by employing generalized likelihood ratio (Harrou et al. 2019) of the human postures. Infact, this reduces the number of video sequence required for training the classifier. The multivariate exponentially weighted moving average method (Harrou et al. 2017) and hidden markov models (Zerrouki and Houacine 2018) are also applied to bypass fall activities from real falls.

Due to the great success in deep learning networks, this is now extended even for the fall detection systems. In some cases, the handcrafted features such as histogram of oriented gradients (Boudouane et al. 2020) and local binary pattern features are augmented with deep learning features to form new hybrid feature set (Wang et al. 2016a, b).

The multi-modal inputs (Khraief et al. 2020) such as RGB video frames, depth images, silhouette of the frames or the stacked optical flow images (Boudouane et al. 2020; Gracewell and Pavalarajan 2019) are fed into Convolutional neural network either for feature extraction or for fall classification (Adhikari et al. 2017; Sehairi et al. 2018). Powerful pre-trained models such as PCANet (Wang et al. 2016a, b), GoogleNet and AlexNet (Chen et al. 2019) are also utilized for fall classification.

The R-CNN, SSDNet and multi-model fully CNN tracker with minimal human intervention are employed to determine the position of humans. In Min et al. (2018), scene analysis is performed with the help of R-CNN to detect the human and furnitures. Later, human falls are detected using the relationship between these furniture and humans in space. Furthermore, the trajectory-weighted deep-convolutional rank pooling descriptor is developed in Zhang et al. (2019a, b) to mitigate the problems due to redundant frames and surrounding environments. In particular, most of the existing human fall detection techniques utilize SVM with non linear kernel function as the classifier, owing to its enhanced generalization capability.

3 Proposed work

The proposed Siamese network based human fall detection framework is depicted in Fig. 1. It consists of a pair of symmetrical convolutional neural networks with inbuilt weight sharing mechanism. Based on the type of convolutional filter used, two different frameworks have been configured, viz., one with 2D standard convolutional filters (2DConv) and the other with Depth wise convolutional filters (DepthConv). In addition, these frameworks are fed with two sets of input features, viz., stacked RGB features and optical flow features. The complete description of the various components present in the proposed framework is provided in the later sections.

3.1 One shot classification

In recent past, one shot classification is considered to be one of the potential research areas in machine learning, due to its ability to learn from constrained datasets and can handle new class of input sequences added to the existing dataset, with which the network has been trained earlier. The network used herein generates the distance metric between a pair of input sequences, rather than the classification score of each class label. During training, the network picks an input sequence as the reference and measure the distance metric with other input sequences of the same class (positive) or other classes (negative).

Some of the differences exhibited by one shot classification from the traditional classifiers are, in case of traditional classifiers an input sequence is fed into sequence of hidden layers and finally, the classification layer computes the probability distribution over all the class labels. However, in one shot classification, softmax based classification layer is replaced with distance metric learning. Further, the traditional approach requires huge number of input sequences on each class and the number of input sequences on each class

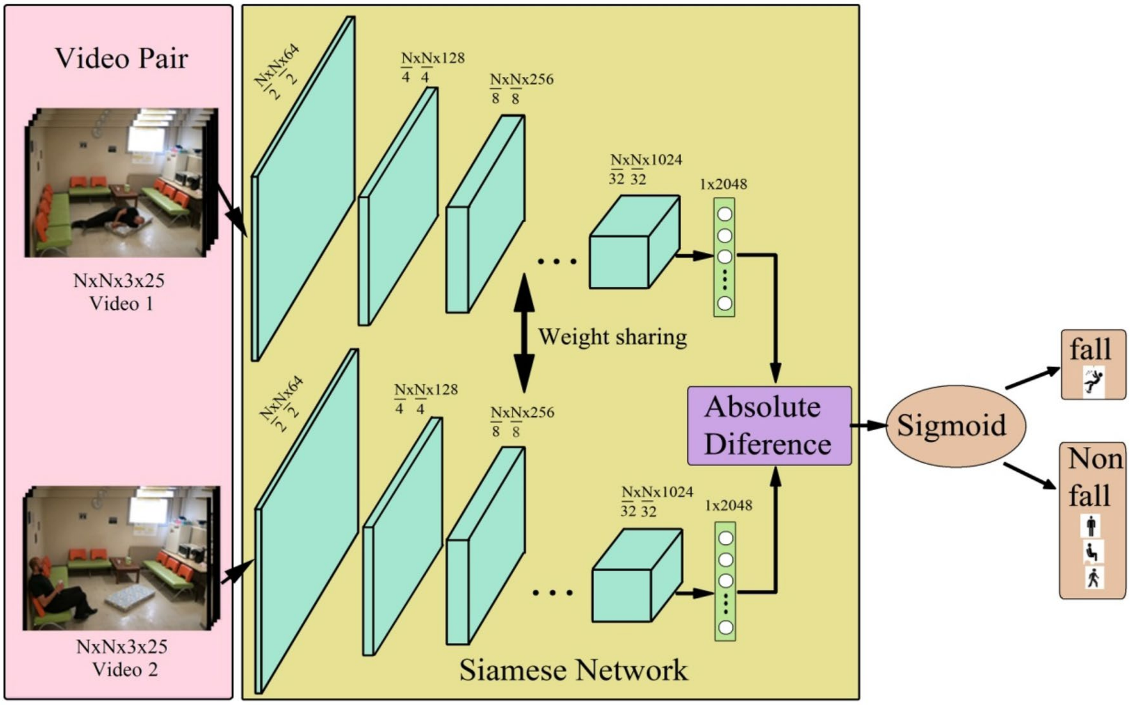


Fig. 1 Proposed architecture for human fall detection

label should be balanced, so as to overcome underfitting and overfitting problems. However, the one shot classification performs well, even for the dataset with many different input classes, but with few input sequences per class (Chopra et al. 2005). In addition, the traditional network has to be typically re-trained from scratch, when new class of input sequences are included to the original dataset. Fascinatingly, in one shot classification, the network can generalize well for the new input class, even after training the network with few input sequences pertaining to the new class.

3.2 Siamese network architecture

The one shot classification is done, utilizing the symmetrical network called Siamese network. Typically, Siamese network is a twin network that contains two symmetrical networks, but they are the copies of the same network. The network receives two distinct inputs, passes through sequence of hidden layers and is merged at the top layer with similarity metric. However, both the networks are assumed to share common weight values through the phenomena called weight sharing. Basically, due to this weight sharing mechanism, it is restricted to generate different feature maps for the two extremely similar video sequences in their respective networks and vice versa. Thereafter, the weighted L_1 distance is computed between the feature vectors generated from the twin networks. At the classification layer, sigmoid based activation function is utilized to map the distance values within the interval [0, 1], where 1 signifies complete similarity and 0 signifies no similarity.

Consider a pair of video sequence X^1 and X^2 is fed into Siamese network made with ‘L’ layers of symmetrical convolutional neural networks. Each convolutional layer employs RELU activation function followed by batch normalization and generates the feature maps $g^{1,k}$ and $g^{1,k}$ at every kth layer. The output of the final convolutional layer is flattened to form a one dimensional vector and is fed into a fully connected layer. Thereafter, the element wise similarity between the feature vectors are computed using L_1 distance metric. It is then subjected to sigmoid based activation unit to generate the probability of occurrence of each class as shown below

$$p(X^1, X^2) = \Phi\left(\sum_{i=1}^{N} \Omega_i \left| g_i^{1,L-1} - g_i^{2,L-1} \right| \right) \qquad (1)$$

Here, N represents the number of features generated by the last fully connected layer, Ω_i is the weighting parameter which is learned during training and Φ is the sigmoid function and is expressed as given below

$$\Phi(z) = 1/(1 + exp\,(-z)) \qquad (2)$$

Since, the Siamese network is trained based on the distance metric, the hypothesis is generated in such a way that the feature vectors generated by the network are different when video sequences of different class is presented into the network and are same when video sequences of same class is presented into the network. The number of convolutional filters used in the successive layers is doubled starting from the first convolutional layer to the final fully connected layer.

3.3 Convolutional filters

Two different convolutional filters, viz., standard convolutional filter and depthwise separable convolutional filter is used to analyze the Siamese Network. Figure 2 depicts the difference between the standard 2D convolutional filter (2DConv) and depthwise separable convolutional filter (DepthConv). Different from a standard convolution, wherein both convolution and merging of their outputs across the channels are computed in a single step, the depthwise separable convolution factorizes the standard convolution into a depthwise convolution and a pointwise convolution. The depthwise convolution initially applies a set of spatial filters of dimension K × K to filter each input channel, followed by this, a pointwise convolution is performed across the channels to combine the outputs of the depthwise convolution. This splitting of convolution is considered to have reduced computational cost and model size (Howard et al. 2017).

In the standard convolution, the 3-D convolutional kernel (ψ) of size (K × K × P) is convolved with the input feature map 'I', to produce the input feature map of a given channel size $N_I \times N_I$. Here, K×K is the spatial dimension of the kernel and P is the number of input channels. The formula to compute the feature map corresponding to the qth channel is given by

$$G_{m,n,q} = \sum_{i=1}^{N_I}\sum_{j=1}^{N_I}\sum_{p=1}^{P} \psi_{i,j,p,q} I_{m+i-1,n+j-1,p} \qquad q = 1,2,\ldots,Q \tag{3}$$

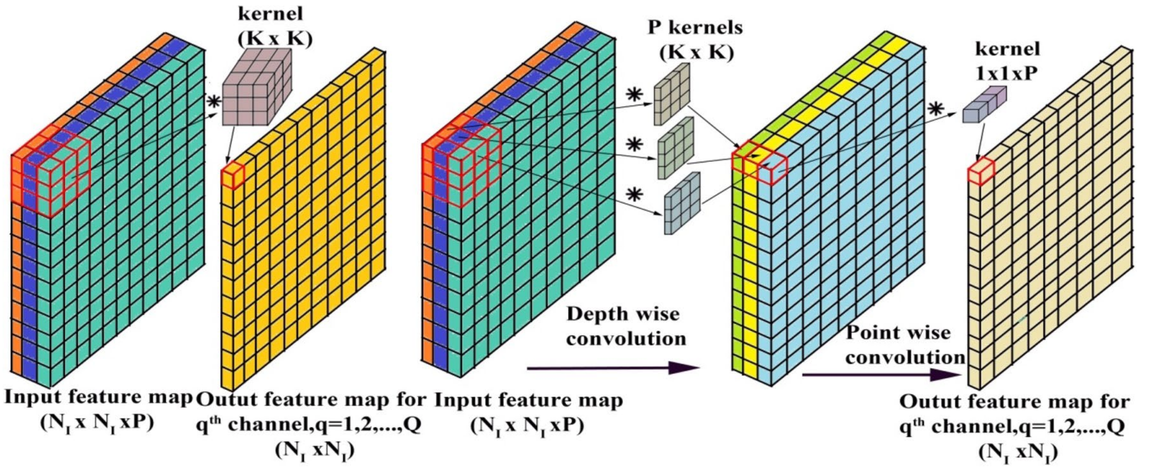
Here, Q is the number of output channels, (m,n) is the spatial location and $N_I \times N_I$ is the spatial dimension of the output feature map in each channel. The computational cost of standard convolution is $\mathrm{K} \times \mathrm{K} \times \mathrm{P} \times N_I \times N_I \times Q$.

In depthwise separable convolution, initially, P number of K × K spatial kernels are used to produce P number of intermediate feature maps. The feature vector of the pth channel is generated using the equation

$$\widehat{B}_{m,n,p} = \sum_{i=1}^{N_I}\sum_{j=1}^{N_I} \widehat{\psi}_{i,j,p} I_{m+i-1,n+j-1,p} \qquad p = 1, 2, \ldots, P \tag{4}$$

Here, ($\widehat{\psi}$) represents the 2-D convolutional kernel and 'I' represents the input feature map. Followed by this is a point wise convolution, in which, these feature maps are stacked and $(1 \times 1 \times \mathrm{P})$ kernel is applied along temporal domain to combine the output of depthwise convolution. The number of multiplications in the first stage is $(K \times K \times N_I \times N_I \times P)$ and the second stage will have a computational cost of $(N_I \times N_I \times P \times Q)$. Thus, the depthwise separable convolution has the computational cost of $\mathrm{K} \times \mathrm{K} \times \mathrm{P} \times N_I \times N_I + \mathrm{P} \times \mathrm{Q} \times N_I \times N_I$. Thereby, this splitting of convolution reduces the computational cost of depthwise separable convolutional network as compared to standard convolutional network by a factor given by

$$\frac{\mathrm{K} \times \mathrm{K} \times \mathrm{P} \times N_I \times N_I + \mathrm{P} \times \mathrm{Q} \times N_I \times N_I}{K \times K \times N_I \times N_I \times P \times \mathrm{Q}} = \frac{1}{\mathrm{Q}} + \frac{1}{\mathrm{K}^2} \tag{5}$$



(a) (b)

Fig. 2 **a** Standard 2D convolution. **b** Depthwise separable convolution.

3.4 Classification in Siamese network

The Siamese network (Zhang et al. 2020) employed in the proposed work utilizes one shot learning mechanism to train the network. The network accepts a pair of video sequences as the input and performs learning, based on the similarity distance computed between the given pair of video sequences. Typically, the proposed framework is developed for binary classification problem, therefore the binary cross entropy based loss function is utilized at the final layer.

Consider a Siamese network receiving a pair of video sequences, denoted by, $X_i = (X_i^1, X_i^2)$. Let 'y' denote the similarity score between the sequences constituting the ith pair, which is denoted by

$$y = \begin{cases} 1 & \text{if the sequences belong to the same class} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The cross entropy for the binary classification problem is defined as

$$\epsilon(y, p) = -[ylog(p) + (1 - y)\log(1 - p)] \tag{7}$$

Here, 'p' denotes the probability that the input video sequences belong to the same class. The loss function used in the proposed architecture is defined as

$$\zeta(X_i) = \epsilon(y, p) + \lambda^T |w|^2 \tag{8}$$

A regularization term (λ) is included to the cross-entropy loss function and finally, the back propagation technique based on gradient descent algorithm is employed to train the network.

After training the proposed framework based on one shot learning mechanism, the network is now ready to generalize a new action sequence. Given a test video sequence X_{test}, it is paired with 'N' number of randomly selected video sequences $\{X_i\}_{i=1}^N$ drawn from the training set. Thereby, each of the 'N' pairs are presented to the network and subsequently, the average similarity score of these 'N' pairs is computed to determine the class label of the given test sequence. Thus, the output class label for the given query video sequence is predicted corresponding to the maximum similarity (P_c) as given below

$$c^* = \text{argmax}_{c=0,1}\{P_c\} \tag{9}$$

4 Experimental results and interpretation

To access the ability/effectiveness of the proposed work, analysis is made on two publically available datasets where action sequences are collected from single calibrated camera. Detailed information of human fall analysis utilizing Siamese network are described in the subsequent sections. The proposed algorithm is implemented in Python on Google Colab, which is the freely available cloud platform.

4.1 Datasets used

The UR fall detection (URFD) dataset (Kwolek and Kepski 2014) contains 30 fall video sequence and 40 other activity sequences acquired by the camera fitted parallel to the floor. Two types of fall are have been pictured here. The first type is 'fall from standing' and the other 'fall from sitting on a chair'. There are other video sequences, which picture other daily activities, viz., walking, squatting, sitting down and picking up an object. Videos are captured at 25 frames/second. The dataset contains intentional falls taken from office setup, performed by five male volunteers more than 26 years of age. Each actor performs three kinds of fall actions, viz., backward, forward and lateral falls, with each action repeated thrice.

FDD (Charfi et al. 2012) contains 250 video sequences with 192 falls and 57 other activity sequences captured with a frame rate of 25 frames/second. The actions are performed by young male volunteers. The set of daily activities (non-fall category) includes walking, crouching down, moving a chair and housekeeping. It includes three types of falls viz., fall due to imbalance, forward falls and falls on improper sitting. Some of the fall sequences in this dataset appear far away from the camera. The video sequences are recorded in different scenarios, Lecture room, Home, Office and Coffee room. Some of the sample image sequences from both the datasets are depicted in Fig. 3.

4.2 Implementation details

Only 20 frames are chosen from the video sequences based on uniform sampling. Two types of Siamese based

Fig. 3 Sample frames from fall/non fall video sequences

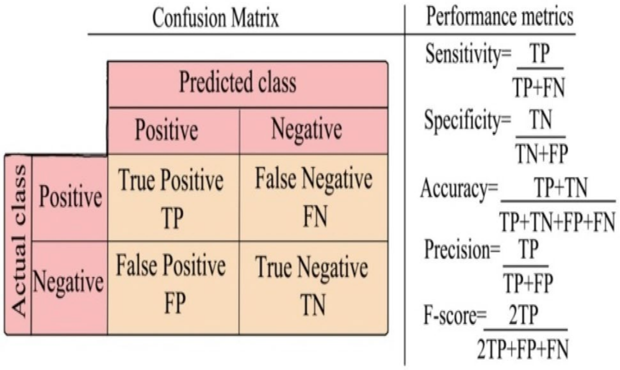
framework is considered, one built with standard convolution (2DConv) filters and the other one with depthwise separable convolution (DepthConv) filters. Further, these two frameworks are operated on two sets of features, viz., stacked RGB frames and stacked optical flow frames generated by PWCNet (Sun et al. 2018; Berlin et al. 2020). The proposed method is validated based on 3-fold cross validation, using the data set mentioned in Sect. 4.1.

4.3 Performance metrics

The proposed frameworks for human fall detection are quantitatively analyzed based on the performance metrics such as accuracy, precision, sensitivity, specificity and F-score. In the confusion matrix, true positive (TP) measures the correct occurrence of fall sequences, false positive (FP) is a false alarm that represents the daily activities misclassified as fall sequence, true negative (TN) indicates the number of correct identification of daily activities and false negative (FN) represents the number of incorrect identification of fall sequences. The evaluation metrics used are listed in Fig. 4.

4.4 Impact on network size

For the proposed human fall detection system, the impact of number of hidden nodes in Siamese network is examined on both URFD and FDD datasets. The performance of the proposed frameworks made with standard convolutional filters (2DConv) and Depth wise separable convolutional filters (DepthConv) are investigated, fed with two sets of inputs viz., RGB frames and optical flow (OF) frames. In both the frameworks the first layer is built with standard 2D convolutional filter. Several experiments are carried out varying the number of hidden layers to identify the effective size of the network. Unlike the number of hidden layers used, the number of feature channels used in the successive layers is doubled (i.e., 64, 128, 256, 512, 1024 and 2048) starting from the first convolutional layer to the final fully connected layer. Followed by this, the final layer of the network is found to be the binary classification layer. For the 4-layered network (L4), two convolutional layers and one fully connected layer that constitute the feature channels of 64, 128 and 256 are used and a classification layer is appended at the end. For the 5-layered network (L5), three convolutional layers and one fully connected layer that constitute the feature channels of 64, 128, 256 and 512 are used. For the 6-layered network (L6), four convolutional layers and one fully connected layer that constitute the feature channels of 64, 128, 256, 512 and 1024 are used. Similarly, for the 7-layered network (L7), five convolutional layers and one fully connected layer that constitute the feature channels of 64, 128, 256, 512, 1024 and 2048 are used. This framework is then followed by L_1 distance computation, and then the probability of occurrence of class labels by using sigmoid activation function. Finally, these configurations are trained using one shot learning mechanism as described in Sect. 3.4.



Confusion Matrix

		Predicted class	
		Positive	Negative
Actual class	Positive	True Positive TP	False Negative FN
	Negative	False Positive FP	True Negative TN

Performance metrics

$$\text{Sensitivity}=\frac{TP}{TP+FN}$$

$$\text{Specificity}=\frac{TN}{TN+FP}$$

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision}=\frac{TP}{TP+FP}$$

$$\text{F-score}=\frac{2TP}{2TP+FP+FN}$$

Fig. 4 Performance metrics used for evaluating the proposed architecture datasets

The experiment is repeated for 30 runs, the resulting mean and standard deviations (std) are computed and represented using the error bars. Figures 5 and 6 shows the impact of network size in terms of accuracy for different values of 'N' on URFD and FDD datasets respectively. For the URFD dataset, corresponding to RGB features, maximum accuracy of 96.57 ± 5.25 % (mean ± std) and 92.5 ± 2.59 % are obtained for the 2DConv and DepthConv based framework respectively on the 5-layered network (L5) constituting N = 15. Corresponding to OF features, maximum accuracy of 100 % and 95 ± 0.46 % are obtained for the 2DConv and DepthConv based framework respectively on the 5-layered network (L5) constituting N = 15.

Similarly, for the FDD dataset, corresponding to RGB features, maximum accuracy of 93.97 ± 3.25 % and 88.76 ± 1.49 % are obtained for the 2DConv and DepthConv based framework respectively on 6-layered network (L6) constituting N = 15. Corresponding to OF features, maximum accuracy of 96.51 ± 5.6 % and 91.71 ± 3.25 % are obtained for the 2DConv and DepthConv based framework respectively for the 6-layered network (L6) constituting N = 15.

4.5 Performance analysis

Inorder to evaluate the proposed architectures, different performance metrics such as sensitivity, specificity, accuracy, precision and F-score are considered. As discussed in the earlier section, maximum accuracy is achieved for N = 15 and the performance metrics obtained to this particular value of 'N' are reported in Tables 1 and 2. As noticed in the table, for both of the considered frameworks, the OF feature produces better performance than the RGB features. In addition, it is evident from these tables that, the standard

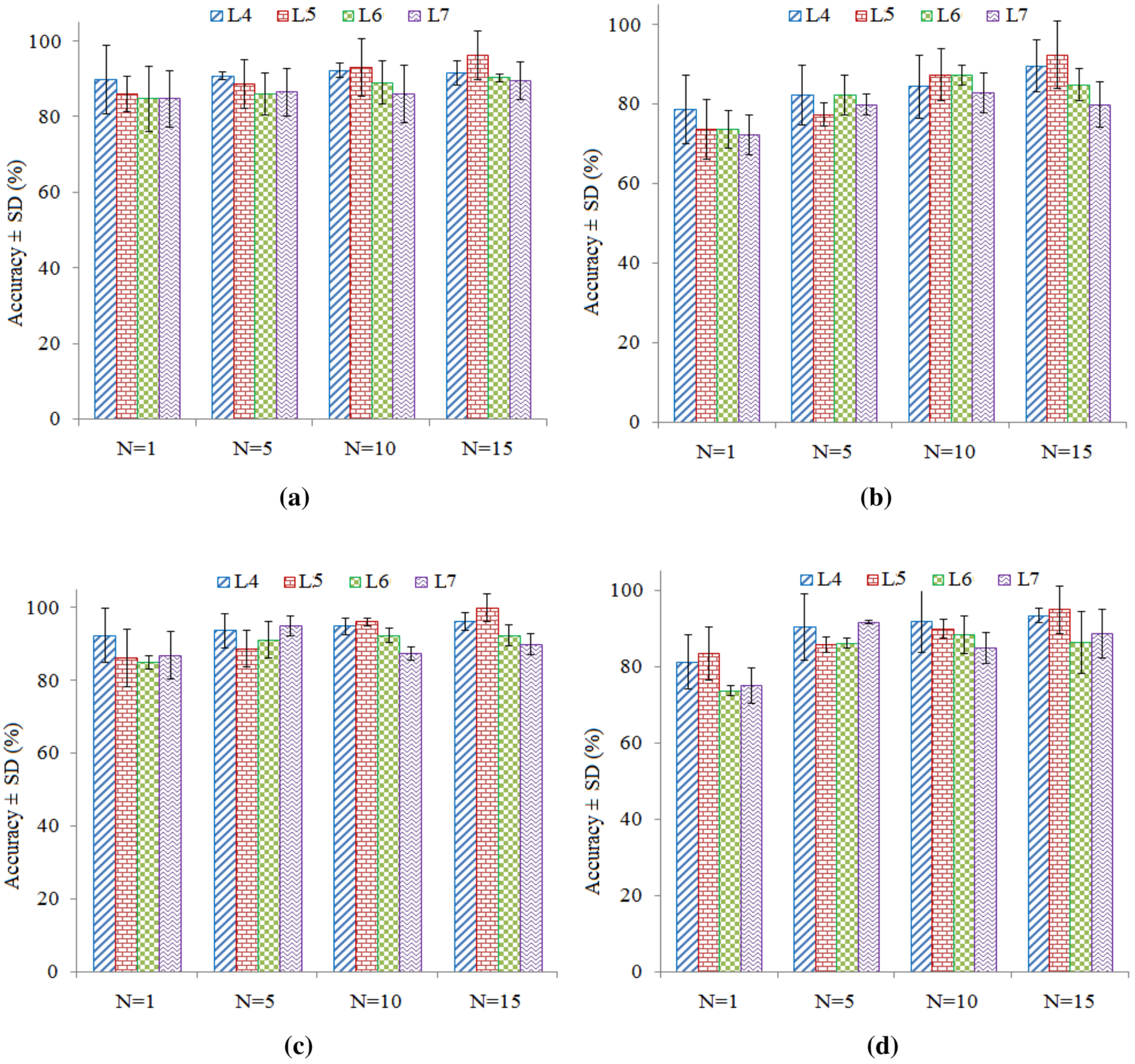


Fig. 5 Impact of network size on URFD. **a** RGB + 2DConv, **b** RGB + DepthConv, **c** OF + 2DConv, **d** OF + DepthConv

convolutional network based architecture (2DConv) consistently performs better than the one based on depthwise separable convolutional network (DepthConv). Therefore, it is concluded that 2DConv based framework operated on OF features performs better than all the other configurations, achieving an overall accuracy of 100 and 97% on the URFD and FDD data sets respectively. But, to extract OF features, it needs a separate module so that it adds some computational burden. Considering on model size, the 7-layered network (L7) consumes 1.4 Mb and 0.82 Mb for 2DConv and DepthConv respectively.

The execution times are 21 and 9 s (for testing) for 2D Conv and DepthConv respectively, when implemented using Python on Google Colab. For the real time implementation of the proposed algorithm, edge processing boards, which are capable of executing machine learning applications at the camera end, viz., Google Coral board, NVIDIA Jetson boards and Intel Neural Compute Stick powered boards may be considered.

As per the results obtained, 100% accuracy is achieved with URFD dataset whereas only 97% is achieved with FDD. The results could be attributed to the fact that, URFD data contains two well defined fall scenarios, one from standing position and the other sitting on a chair and falling from the sitting position. Further, the recordings have been carried out in a controlled set-up without occlusion and with less illumination changes.

However, in FDD, the video sets portray a more realistic environment. The intra-class variability of FDD data set is much higher as the sets very well capture occlusion scenario, variation in illumination, textured background, shadows and reflections. Therefore, the proposed technique has achieved lower accuracy (97 %) on FDD dataset as compared to that on URFD.

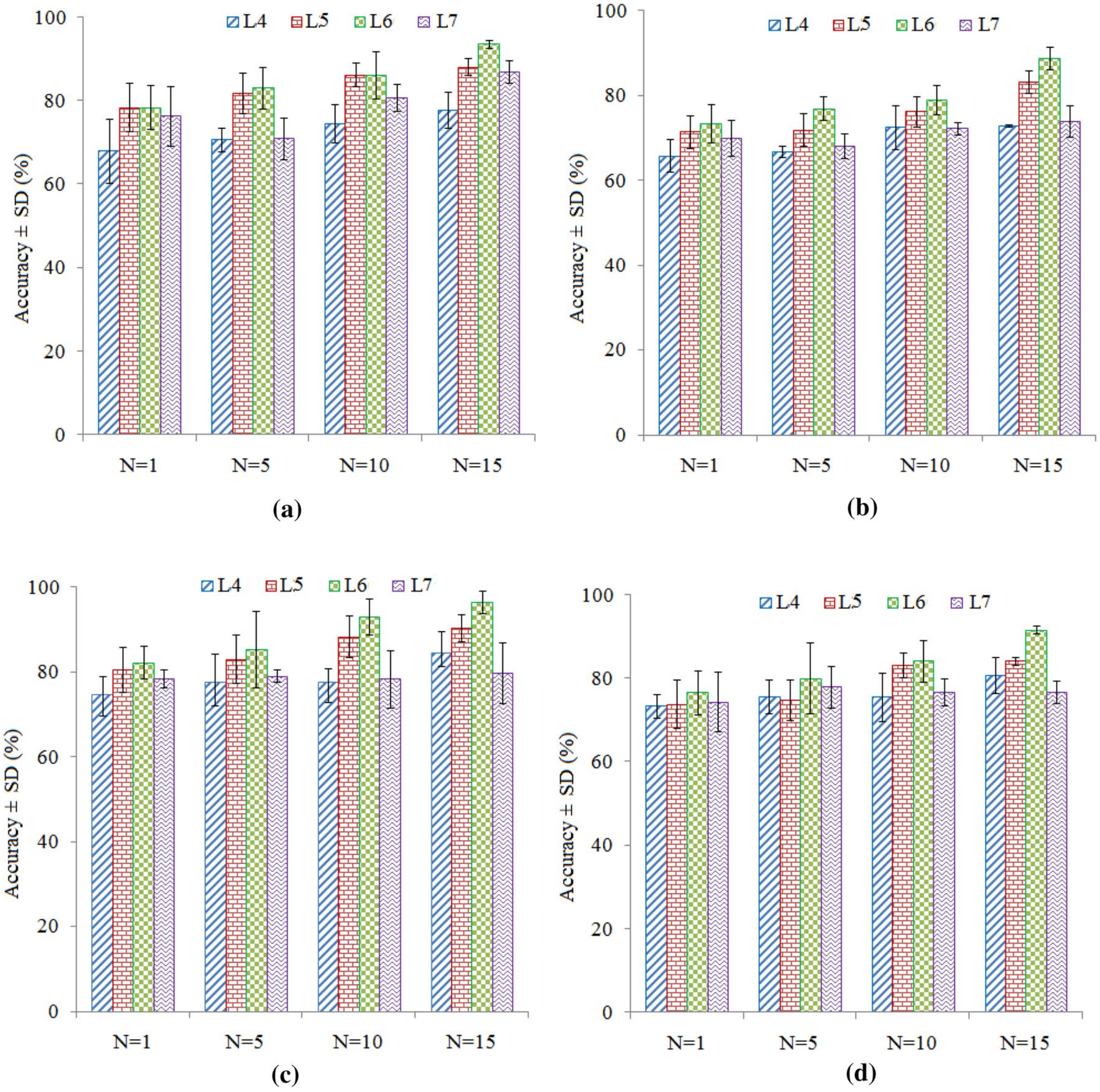


Fig. 6 Impact of network size on FDD. **a** RGB + 2DConv, **b** RGB + DepthConv, **c** OF + 2DConv, **d** OF + DepthConv

Table 1 Performance analysis of Siamese network based human fall detection on URFD dataset

Configurations	Sensitivity (%)	Specificity (%)	Accuracy (%)	Precision (%)	F-score (%)
RGB + 2DConv	87 ± 1	97 ± 6	96 ± 7	96 ± 6	91 ± 8
RGB + DepthConv	89 ± 3	91 ± 1	93 ± 5	95 ± 1	87 ± 6
OF + 2DConv	100	100	100	100	100
OF + DepthConv	93 ± 5	100	97 ± 6	100	96 ± 6

Table 2 Performance analysis of Siamese network based human fall detection on FDD dataset

Configurations	Sensitivity (%)	Specificity (%)	Accuracy (%)	Precision (%)	F-score (%)
RGB + 2DConv	96 ± 3	89 ± 5	93 ± 5	91 ± 7	93 ± 4
RGB + DepthConv	95 ± 4	91 ± 4	88 ± 1	97 ± 2	95 ± 2
OF + 2DConv	97 ± 3	96 ± 2	96 ± 1	96 ± 1	96 ± 1
OF + DepthConv	93 ± 5	87 ± 4	91 ± 2	88 ± 4	90 ± 2

Table 3 Comparison to the state-of-the-art methods in terms of accuracy

Methods	URFD (%)	FDD (%)
Nunez-Marcos et al. (2017)	95	97
Zerrouki et al. (2018)	97	97
Harrou et al. 2019	96	96
Proposed (OF + 2DConv)	100	97

All the video sequences derived from the datasets, viz., URFD and FDD portray young adults performing fall actions and other daily activities from indoor environment such as home and office. However, the intentional falls recorded in these datasets are different from real falls, in terms of speed, spontaneity, and nature of fall (Khan et al. 2017). Therefore, though, the human fall detection system proposed herein performs well with the benchmark datasets, there is likely to be a marginal reduction in accuracy when tested on a typical elderly fall action. However, as Siamese network is capable of performing well with limited training data set, by fine tuning the network with a few real time data samples involving elderly fall action, this problem could be alleviated.

4.6 Comparison with the state-of-the-art methods

Experiments have been carried out on the proposed Siamese network based architecture and compared with that of the other best performing schemes reported so far in terms of accuracy, as presented in Table 3.

In the human fall detection scheme used by Zerrouki et al. (2018) a background subtraction technique is used followed by curvelet based feature extraction. SVM is used for posture estimation and finally a HMM based model is used for action classification. In the method proposed by Harrou et al. (2019), the silhouette is extracted and the feature vectors generated from the silhouette are fed to a Generalized Likelihood Ratio (GLR) based classifier. Further, a SVM based classifier is used to classify between a 'true-fall' and a 'fall like action'. In both these techniques, viz., Harrou et al. (2019) and Zerrouki et al. (2018), foreground is extracted based on background subtraction. In Zerrouki et al. (2018), the background extraction technique used is based on successive frame difference which is very simple. A more sophisticated background extraction technique would be required in a scenario with complex background, which varies dynamically. However, the proposed algorithm is convolutional neural network based, which is capable of carrying out foreground extraction even in a complex background scenario, alleviating the requirement for a separate background extraction module. Therefore, it is not possible to make a direct comparison of the computational complexity of the proposed technique with these two techniques.

The human fall detection system proposed by Nunez-Marcos et al. (2017) is CNN based. The network is initially trained with ImageNet dataset and then through transfer learning, the network is fine tuned for human fall detection with stacked optical flow as the feature set. The proposed technique is based on Siamese network, which is a twin network and therefore requires twice the complexity as compared to the single CNN based technique. However, the advantage of the proposed Siamese network is its ability to learn even with a limited training set.

From the table, it is evident that the proposed framework for human fall detection is comparable with the other approaches on FDD and shows marginal improvement over the other state-of-the-art methods reported on URFD dataset.

5 Conclusions

In the proposed work, efficient deep learning based Siamese frameworks, for human fall detection have been formulated. Two different Siamese based frameworks have been configured, one embedded with standard 2D convolutional filter and the other with depth wise separable convolutional filter. Further, these frameworks are fed either with RGB features or with optical flow features. Though the depthwise separable convolution is computationally simple, it has been experimentally demonstrated that there is a compromise in terms of accuracy as compared to the standard convolutional network. Moreover, it is found that the Siamese configuration operated on optical flow based features outperforms RGB features, as the motion information is efficiently represented by optical flow features. In future, the proposed configuration could be extended to the multi-camera scenario. Another challenging future direction is to extend this to outdoor environments, through the incorporation of additional modules such as region proposal techniques (Yao et al. 2020). Further, edge implementation of the proposed algorithm using hardware boards that can support deep learning architecture could also be considered.

Acknowledgements The first author is a recipient of DST INSPIRE Fellowship and wishes to thank DST, India for the same.

Compliance with ethical standards

Conflict of interest The authors declare that there are no conflicts of interest in the authorship or publication of this paper.

References

Abobakr A, Hossny M, Nahavandi S (2018) A skeleton-free fall detection system from depth images using random decision forest. IEEE Syst J 12(3):2994–3005

Adhikari K, Bouchachia H, Nait-Charif H (2017) Activity recognition for indoor fall detection using convolutional neural network, In: 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), Nagoya, pp 81–84

Berlin SJ, Mala J (2020) Light weight convolutional models with spiking neural network based human action recognition. J Intell Fuzzy Syst 39:961–973

Bhavya KR, Park J, Park H, Kim H, Paik J (2016) Fall detection using motion estimation and accumulated image map. In: 2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), Seoul, pp 1–2

Boudouane I, Makhlouf A, Harkat MA et al (2020) Fall detection system with portable camera. J Ambient Intell Humaniz Comput 11:2647–2659

Charfi I, Miteran J, Dubois J, Atri M, Tourki R (2012) Definition and performance evaluation of a robust SVM based fall detection solution. In: 8th international conference on signal image technology and internet based systems, pp 218–224

Chen L, Kong X, Tomiyama H, Meng L (2019) Multiple states fall detection system for senior citizens. In: Proceedings of the international conference on advanced mechatronic systems (ICAMechS), Japan, pp 169–174

Chopra S, Hadsell R, LeCun Y (2005) Learning a similarity metric discriminatively, with application to face verification. In: IEEE computer society conference on computer vision and pattern recognition (CVPR'05), USA, pp 539–546

Deandrea S, Lucenteforte E, Bravi F, Foschi R, La Vecchia C, Negri E (2010) Risk factors for falls in community-dwelling older people: a systematic review and meta-analysis. Epidemiology 21(5):658–668

Fan K, Wang P, Zhuang S (2019) Human fall detection using slow feature analysis. Multimed Tools Appl 78(7):9101–9128

Gracewell JJ, Pavalarajan S (2019) Fall detection based on posture classification for smart home environment. J Ambient Intell Hum Comput

Harrou F, Zerrouki N, Sun Y, Houacine A (2017) Vision-based fall detection system for improving safety of elderly people. IEEE Instrum Meas Mag 20(6):49–55

Harrou F, Zerrouki N, Sun Y, Houacine A (2019) An integrated vision-based approach for efficient human fall detection in a home environment. IEEE Access 7:114966–114974

Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861

Iazzi A, Rziza M, Thami ROH (2020) Efficient fall activity recognition by combining shape and motion features. Comput Vis Media 6(3):247–263

Jansi R, Amutha R (2020) Detection of fall for the elderly in an indoor environment using a tri-axial accelerometer and Kinect depth data. Multidimens Syst Signal Process 31(4):1207–1225

Joshi NB, Nalbalwar SL (2017) A fall detection and alert system for an elderly using computer vision and Internet of Things. In: 2017 2nd IEEE international conference on recent trends in electronics, information and communication technology (RTEICT), Bangalore, pp 1276–1281

Juang C, Chang C (2007) Human body posture classification by a neural fuzzy network and home care system application. IEEE Trans Syst Man Cybern Part A Syst Hum 37(6):984–994

Kerdjidj O, Ramzan N, Ghanem K et al (2020) Fall detection and human activity classification using wearable sensors and compressed sensing. J Ambient Intell Human Comput 11:349–361. https://doi.org/10.1007/s12652-019-01214-4

Khan SS, Hoey J (2017) Review of fall detection techniques: a data availability perspective. Med Eng Phys 39:12–22

Khraief C, Benzarti F, Amiri H (2020) Elderly fall detection based on multi-stream deep convolutional networks. Multimed Tools Appl 1–24

Kwolek B, Kepski M (2014) Human fall detection on embedded platform using depth maps and wireless accelerometer. Comput Methods Programs Biomed 117(3):489–501

Leal-Taixe L, Canton-Ferrer C, Schindler K (2016) Learning by tracking: Siamese CNN for robust target association. In: IEEE conference on computer vision and pattern recognition workshops, pp 33–40

Lotfi A, Albawendi S, Powell H, Appiah K, Langensiepen C (2018) Supporting independent living for older adults; employing a visual based fall detection through analysing the motion and shape of the human body. IEEE Access 6:70272–70282

Liu C, Lee C, Lin P (2010) A fall detection system using k-nearest neighbor classifier. Expert Syst Appl 37:7174–7181

Ma X, Wang H, Xue B, Zhou M, Ji B, Li Y (2014) Depth-based human fall detection via shape features and improved extreme learning machine. IEEE J Biomed Health Inf 18(6):1915–1922

Mastorakis G, Makris D (2014) Fall detection system using Kinect's infrared sensor. J Real-Time Image Proc 9:635–646

Min W, Cui H, Rao H, Li Z, Yao L (2018) Detection of human falls on furniture using scene analysis based on deep learning and activity characteristics. IEEE Access 6:9324–9335

Mirmahboub B, Samavi S, Karimi N, Shirani S (2013) Automatic monocular system for human fall detection based on variations in Silhouette Area. IEEE Trans Biomed Eng 60(2):427–436

Nunez-Marcos A, Azkune G, Arganda-Carreras I (2017) Vision-based fall detection with convolutional neural networks. Wirel Commun Mobile Comput

Ozcan K, Velipasalar S, Varshney PK (2017) Autonomous fall detection with wearable cameras by using relative entropy distance measure. IEEE Trans Hum Mach Syst 47(1):31–39

Rougier C, Meunier J, St-Arnaud A, Rousseau J (2011) Robust video surveillance for fall detection based on human shape deformation. IEEE Trans Circuits Syst Video Technol 21(5):611–622

Rougier C, Meunier J, St-Arnaud A, Rousseau J (2013) 3D head tracking for fall detection using a single calibrated camera. Image Vis Comput 31246–31254

Rubenstein LZ (2006) Falls in older people: epidemiology, risk factors and strategies for prevention. Age Ageing (35)

Ruiz V, Linares I, Sanchez A, Velez JF (2020) Off-line handwritten signature verification using compositional synthetic generation of signatures and Siamese Neural Networks. Neurocomputing 374:30–41

Sehairi K, Chouireb F, Meunier J (2018) Elderly fall detection system based on multiple shape features and motion analysis. In: International conference on intelligent systems and computer vision (ISCV), pp 1–8

Shieh WY, Huang JC (2012) Falling-incident detection and throughput enhancement in a multi-camera video-surveillance system. Med Eng Phys 34(7):954–963

Soni PK, Choudhary A (2018) Automated fall detection using computer vision. In: Tiwary U (eds) Intelligent human computer interaction. IHCI 2018. Lecture Notes in Computer Science, vol 11278. Springer, Cham

Sun D, Yang X, Liu MY, Kautz J (2018) PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: IEEE/CVF international conference on computer vision and pattern recognition, USA, pp 8934–8943

Taigman Y, Yang M, Ranzato MA, Wolf L (2014) Deepface: Closing the gap to human-level performance in face verification. In: IEEE conference on computer vision and pattern recognition, pp 1701–1708

Vallabh P, Malekian R (2018) Fall detection monitoring systems: a comprehensive review. J Ambient Intell Human Comput 9:1809–1833

Wang K, Cao G, Meng D, Chen W, Cao W (2016a) Automatic fall detection of human in video using combination of features, In: 2016a IEEE international conference on bioinformatics and biomedicine (BIBM), Shenzhen, pp 1228–1233

Wang S, Chen L, Zhou Z et al (2016b) Human fall detection in surveillance video based on PCANet. Multimed Tools Appl 75:11603–11613

Wang F, Yang B, Li J, Hu X, Ji Z (2020) Attention-based siamese region proposals network for visual tracking. IEEE Access 8:86595–86607

Yacchirema D, de Puga JS, Palau C, Esteve M (2019) Fall detection system for elderly people using IoT and ensemble machine learning algorithm. Pers Ubiquit Comput 23(5):801–817

Yao C, Hu J, Min W, Deng Z, Zou S, Min W (2020) A novel real-time fall detection method based on head segmentation and convolutional neural network. J Real-Time Image Proc 17:1939–1949

Yu M, Rhuma A, Naqvi SM, Wang L, Chambers J (2012) A posture recognition-based fall detection system for monitoring an elderly person in a smart home environment. IEEE Trans Inf Technol Biomed 16(6):1274–1286

Zerrouki N, Houacine (2018) A Combined curvelets and hidden Markov models for human fall detection. Multimed Tools Appl 77:6405–6424

Zhang Z, Ma X, Wu H, Li Y (2019a) Fall detection in videos with trajectory-weighted deep-convolutional rank-pooling descriptor. IEEE Access 7:4135–4144

Zhang Z, Zhang Y, Cheng X, Li K (2019b) Siamese network for real-time tracking with action-selection. J Real-Time Image Process 1–11

Zhang C, Wang H, Wen J, Peng L (2020) Deeper siamese network with stronger feature representation for visual tracking. IEEE Access 8:119094–119104