Fall Detection Based on RetinaNet and MobileNet Convolutional Neural Networks

2020 15th International Conference on Computer Engineering and Systems (ICCES) | 978-0-7381-0559-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCES51560.2020.9334570

Hadir Abdo
Information Technology Department
Faculty of Computers and Information
Menoufia University, Egypt
hadir.itunit@ci.menofia.edu.eg

Khaled M. Amin
Information Technology Department
Faculty of Computers and Information
Menoufia University, Egypt
k.amin@ci.menofia.edu.eg

Ahmad M. Hamad
Information Technology Department
Faculty of Computers and Information
Menoufia University, Egypt
ahmahit@ci.menofia.edu.eg

***Abstract*—The problem of falling is a major health problem resulting in serious injuries and sometimes lead to death especially for elderly. Elderly people aged over than 75 are exposed to accidental deaths due to falls. Approaches based on computer vision give a promising and an effective solution for detection human falls. This paper presented a method for fall detection which based on combining convolutional neural networks RetinaNet and Mobilenet in addition to handcrafted features. Traditional human detection methods may result in human shape deformation which affect the performance of fall detection frameworks. Therefore, the proposed framework depends on RetinaNet for detecting humans with shorter computing time and higher accuracy compared with the traditional human detection methods. Then, the proposed framework relies on handcrafted features to represent shape and motion properties of the detected human. The proposed framework extracts aspect ratio and head position as shape features and motion history image as a motion feature of the detected human to create the feature map. This feature map is used in training MobileNet network to classify the human motion into fall or not-fall. The proposed framework is evaluated using UR and FDD datasets and the experimental results proved the efficiency of the proposed framework achieving up to 98% accuracy compared with the state-of-the-art methods.**

***Keywords*—human fall detection; vision fall detection CNN for fall detection**

I. INTRODUCTION

The safety and rescue of elderly people and patients from death and injuries become an urgent goal of computer vision and modern technologies. Elderly people are often not able to get up after their fall. Therefore, elderly are more likely to die accidentally because of the falling [1]. At present, many automatic fall detection systems are used to monitor elderly and sick people in homes and hospitals.

Fall detection systems can be classified into: ambient devices, wearable devices and vision-based devices (cameras). The wearable-based detection includes the usage of devices such as accelerometers and gyroscopes [2]. Although these devices are easy to use, low prices, and can help in monitoring persons everywhere, but such devices suffer from disadvantages such as user forgetting, easy to theft, or forgetting to recharge the empty battery. These drawbacks make such devices an inconvenient option for elderly people.

Ambience based approach relies on collecting instantaneous data of the active targeted regions around the elderly persons by using external sensors. Pressure and vibration sensors are examples of ambient sensors which are planted on the floor to indicate the instantaneous location of the persons. Although this approach is cheap and non-intrusive, but such sensors cause a high rate of false alarms because of its sensitivity to everything in the active area leading to low detection rate.

Due to the spread of cameras and the advances of computer vision, fall detection-based vision approaches present a promising solution for caring elderly people. Vision based systems are best suited and characterized over the previous systems in localizing and monitoring multiple people besides detecting multiple actions in the same time. In addition, the elders don't have to carry specialized devices. A camera is easy installed and gives an immense amount of information about the people such as location, actions, and motion [3]. In this approach, the moving object is detected, classified as a human, and tracked in each frame of the video. Using computer vision techniques, fall identifying depending on useful features which are extracted and analyzed for the detected person such as extracted silhouettes. [4]. Recently, deep learning is presented as a powerful technique in various problems of computer vision. Using deep learning in detecting the fall of elderly people improves the fall-detection accuracy and reduces the false alarms.

This paper is organized as follows: Section 2 describes related works fall detection systems. In Section 3 the methodology and proposed method. In Section 4 experimental setup and the results. In section 5 conclusions and discussions

II. RELATED WORK

The spread of cameras everywhere and the un necessity to take any equipment along make vision-based methods the most appropriate to detect falls automatically. The steps of most vision fall detection systems include video acquisition, video analysis, and alarming communication. The performance of vision- fall detection systems depends mainly on the analysis of video frames to detect falls. The most difficult challenge of video analysis step is the deformation of the human shape due to poor detection results caused by many reasons such as moving background, illumination changes, etc. Other challenges like occlusion, camera position, and the similarity of human shape in different actions (like sitting and falling) may also result in poor performance of fall detection approaches.

Mixture of motion gradients and human shape variation

features were presented by M. Khan et al [5] to detect human fall. Large movements is used to characterize human falling. Tracking the person motion is used to determine global motion orientation which decides the direction of the human movement. The fall is distinguished from other activities, depending on the aspect ratio which is tested against a specific threshold. False alarm may be produced in the case of crouching down action due to the similarity of aspect ratio between fall and couching down actions.

Another fall detection approach based on the human posture recognition depending on human silhouette analysis is presented by Yu et al[6]. In order to fade the problem of wrong analysis for the human shape, the background subtraction technique is used. Then, the description of different postures is performed by using a projection histogram and a fitting ellipse as global and local features. Finally, by using graph support vector machine (DAGSVM), the fall is classified. This approach suffers from two challenges. The first challenge is the problem of occlusions and the secobd challenge is the problem of the extistence of multiple moving objects in the scene. Shukla and Tiwari [7] proposed an approach for detecting human fall that is based on two features. These two features are the height of the center of the human body to the ground and the history of human movement. The approach detects the object by applying the background subtraction which gives a complete silhouette of the moving object. Then, the noise is removed to track the object correctly. The classification is performed by determining the centroid of the object and MHI. The fall classified according to a threshold angle as If the center point of the silhouette has value less than the specific threshold; then the movement is a fall, otherwise it is no fall.

Zerrouki et al. [8] presented a fall detection approach based on a human silhouette shape variation in vision monitoring. This system adapts HMM and SVM for identifying fall and non-fall in video relying on the information extracted from the RGB camera which leads to a misclassification of some falls especially in case of low quality images. It is also difficult to extract the silhouette of a human body using a traditional RGB camera, especially in cases of darkness or occlusion.

Doulamis and N. Doulamis [9] introduced an adaptive deep learning approach to detecting human falls. First, based on the supervised learning approach, it isolates the introduction from the background. After that, the network is trained in human body models that describe the human body as a place under the human face, and this is achieved using a two-dimensional capability that leads to pixels that belong to the human body in relation to its location in the region. Finally, we rely on a decision-making mechanism that allows network weight to be adjusted to suit current visual conditions while a decision-making mechanism makes sure that the current environment is ”similar” to acquired knowledge or not, accordingly the human fall is categorized. False alarm occurs because of a lack of activity in the movement or the presence of various movements, and thus the movement of human is ignored.

Marcos and G. Azkune [10] presented a fall detection approach as it classify the presence of a fall in a sequence of frames depending on CNN (convolutional neural networks). A set of optical flow images were used and the features were extracted and classified using VGG-16 CNN model which pre-trained on ImageNet. Although this approach has given high accuracy and efficiency, but it suffers from event misclassification due to the lack of enough training network.

To cope with previously mentioned challenges, the proposed framework presented in this paper combines handcrafted features and onvolutional neural networks RetinaNet and Mobilenet. RetinaNet has the highest accuracy among convolutional neural networks used for human detection. In addition, Retinanet copes with challenges of traditional background subtraction techniques. Therefore, the proposed framework depends on RetinaNet to detect the location of the human in a sequence of frames accurately. Then, bounding box (surrounds the detected human) aspect ratio and the head position are detected as a representation for the detected human shape features. Furthermore, motion history image (MHI) which aggregates the motion of a moving object in a single image is obtained. These features are used to form the feature map that used to train Mobilenet network which is used to classify the human action to fall or not-fall. MobileNet uses two global hyper-parameters that efficiently trades off between accuracy and latency. The proposed framework exploits the efficiency and the accuracy of the convolutional neural networks RetinaNet and Mobilenet. To assess the efficiency of the proposed framework, the proposed framework is evaluated using UR and FDD datasets and the experimental results proved the efficiency of the proposed framework achieving up to 98% accuracy compared with the state-of-the-art methods.

III. PROPOSED METHOD

The proposed fall detection framework presented in this paper depends on CNN networks for detecting moving humans in videos and classifying the human motion to ”fall” or ”not fall”. The three phases of the proposed framework are shown in Fig.1. The first phase of the proposed framework is to detect and track moving humans in video frames. RetinaNet is the selected model to detect moving humans. RetinaNet showed a high level of accuracy with acceptable rate of detection speed [11]. The second phase of the proposed framework is to extract feature of human motion and human body shape. These features include width to height ratio of the bounding box surrounding the human body, motion history images which show the direction of object movement, head tracking, and orientation which indicates of the object position in relation to the ground. These features are tracked in the frame sequence and fed into the third phase to classify human motion ”fall” or ”not fall”. The third phase of proposed method employs a modified version of MobileNets [12] to classify human motion according to the features extracted in the second phase into ”fall” or ”not fall”. MobileNet feeds the first convolution layer with the extracted features and passing results to the following convolutional layers which in turn make a classification of the inputs by applying depth wise convolution until to reach to the last layer which give a decision based on the output of last layer.

A. Human detection

Most previous fall detection methods is based on background subtraction for detecting objects and then classifying these objects to human or non-human. These methods that is based on background subtraction method suffer from challenges like illumination changes, dynamic background,

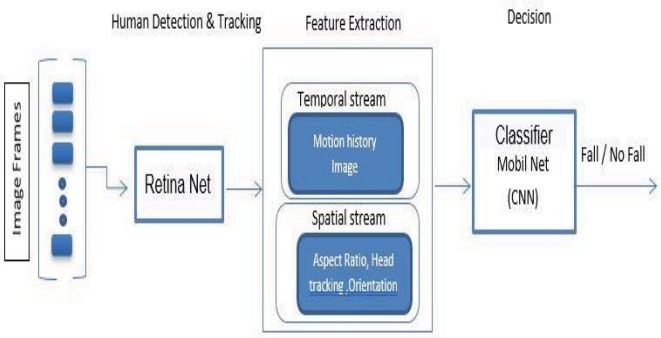

Fig. 1. An overview of the proposed framework phases.

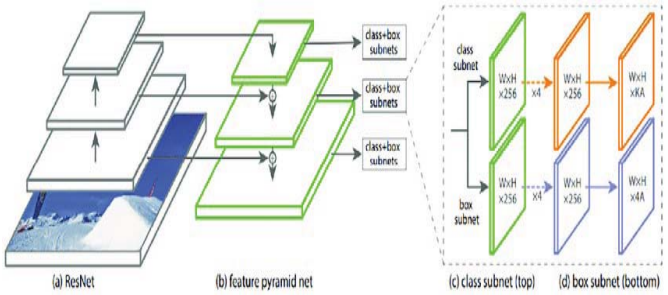

Fig. 2. RetinaNet model architecture.

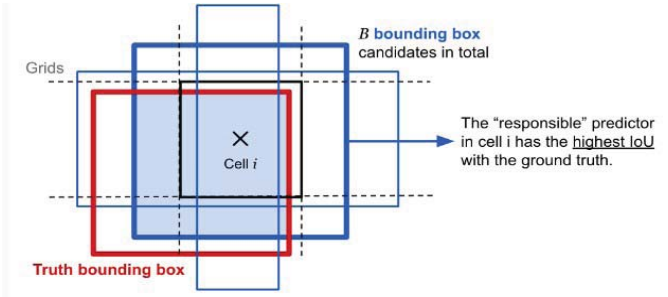

Fig. 3. Bounding box generation.

camouflage, etc. These challenges may result in human shape deformation and therefore affect the performance of fall detection frameworks. Unlike previous methods, the proposed method uses deep convolutional neural network "RetinaNet" as a human detection network. RetinaNet gives the highest accuracy among other convolutional neural networks used for human detection and cope with challenges of traditional background subtraction techniques [13].

Hierarchical structure of RetinaNet includes two subnets and a backbone network. RetinaNet backbone extracts the features and then the two subnet layers classify and regress objects based on the output of the backbone. Fig.2 shows the general architecture of RetinaNet. The first subnet classifies objects according to the backbone's output. The second subnet use the backbone output to construct a bounding box regression. The feature map of input frame is extracted by Feature Pyramid Network (FBN) which is the backbone of the RetinaNet. Anchor boxes similar to those in the RPN are used to identify the bounding of objects at each position. These anchor boxes are modified for multi-class detection with adapted thresholds. These anchors include areas from 32 to 512 on P3 to P7 levels of the pyramid, respectively, and aspect ratios of 1:2, 1:1, and 2:1. Therefore, there are 9 anchors for each level and across the existing levels covering the scale range from 32 to 813 pixels according to the input image to network. As shown in Fig.3, each anchor is set to a one vector and each vector is set to one of K object classes and a 4-vector for box regression. Then, the regression subnet assigns each anchor to a nearby ground truth object boxes according to the intersection-over-union (IoU) with threshold of 0.5 (set empirically). Thereafter, the classification subnet predicts presence or absence object at each position for each of A anchors and K objects classes. If an anchor is unassigned to a ground truth, this anchor is ignored during training. The output of this subnet is the spatial location of the object in the scene as shown in Fig.4.

B. Feature extraction

While the proposed method depends on CNN for human detection and classification for increasing the accuracy, the proposed method extracts the features from the detected object manually instead of automatic feature extraction to decrease the computation time. One of the major issues in fall detection systems is the selection of features to be classified into fall or not-fall. Moving from the initial data space to a feature space that will make the fall detection more recognizable. The proposed method depends on the detected object shape and motion features to decide the fall. To retain rich object information, a width to height ratio for the bounding box(apect ratio), object motion history image, and head tracking features are selected.

1) Aspect Ratio: In object detection, bounding box is used to describe the object location [14]. A bounding box can be defined as a rectangular square that is defined by the x and y axis coordinates in the upper left corner and the x and y axis coordinates in the lower right corner. Aspect ratio denote the ratio between box width and height. It stands to reason that the ratio of width to height in case of fall is large and is small in case of movement or crouching down.

2) Motion History Image (MHI): Motion history image (MHI) is introduced by Bobick and Davis as an action representation approach that aggregates the motion of a moving object in a single image [15]. MHI is a template image that is used for recognizing and understanding the human action as well as the direction of the movement. The aggregation of

Fig. 4. Detection result under IOU threshold =0.5.

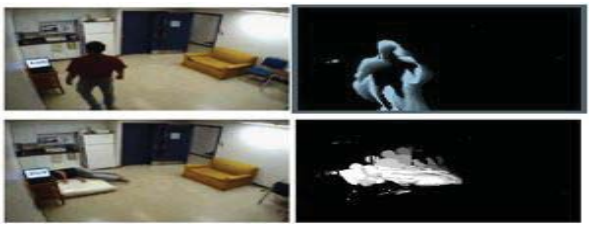

Fig. 5. Example of MHI images for two different actions.

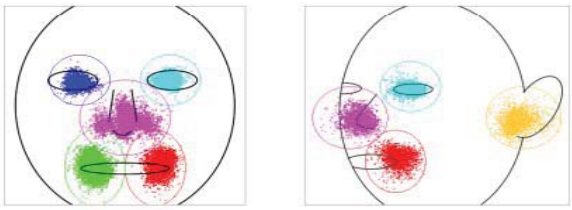

Fig. 6. Example of the facial feature detectors for the frontal and left profile views.

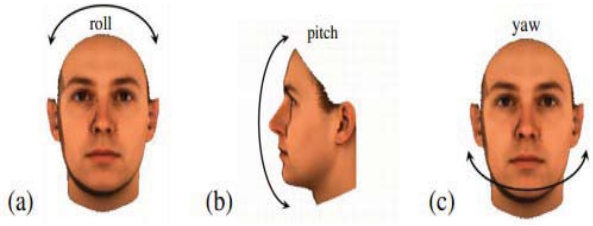

Fig. 7. Three head angles (a) roll angle, (b) pitch angle, and (c) yaw angle.

the human motion as well as the direction of the motion are fed to train the classifier in the following phase. The proposed method depends mainly on the motion aggregation and the motion direction as strong indicator to decide the human fall. The definition of the MHI illustrated by eq.1:

$$H_\tau(x,y,t) = \begin{cases} \tau, & \text{if D(x,y,t)=1.} \\ max(0, H_\tau(x,y,t-1), & \text{otherwise.} \end{cases} \quad (1)$$

Where D(x,y,t) refers to a the value of the pixel (x,y) at time t of a binary image D with '1's refers to the motion regions. The frame differencing method is used to obtain D(x,y,t). $H_\tau(x,y,t)$ is a scalar-valued image in which regions in last motion occurred is brighter than motion in previous frames. Fig.5 shows an example of two motion history images for two different actions.

3) Head tracking: The proposed method depends on detecting and tracking the head of moving human because the head movement during the human fall is noticeable. The head is approximated by an ellipse. The proposed method detects the human head through detecting the human face in the image by applying a face detection system. A boosted cascade of the same Haar-like is exploited to detect faces in frames [16]. As shown in Fig.6, a set of two feature groups are used for detecting faces in both frontal and profile views. The features for detecting faces in frontal view are: left and right eye, nose, left and right mouth corner. The features of the entire visible part of the face are used as the features for detecting faces in profile view. By applying a sliding window on the input frame to search about the frontal or the profile features, a bounding box is positioned for the detected face. Once the face was detected, a neural network is used to estimate the three rotation angles roll, pitch, and yaw of the head pose as shown in Fig.7. A multi-layer feed-forward network with one hidden layer is used with each unit computes its output by passing the incoming weighted signal through an activation function. The input of this network is the coordinate (x ,y) which determines the position of a facial feature and the distance between pairs of facial features *i* and *j* $(d_x = x^i - x^j, d_y = y^i - y^j)$ both these values are normalized in the range $[-0.5, +0.5]$. The angles ranges that the head detector cover are $[-25, +25]$, $[-40, +40]$, and $[-90, +90]$ for pitch, roll, and yaw, respectively.

C. Classification

The proposed fall detection framework presented in this paper depends on a light weight deep convolutional neural network MobileNet. MobileNet is a fast, an accurate, and a small size deep convolutional neural networks that can be used for classification and detection. A modified version of a MobileNet network is used for classifying extracted features at the previous phase and gives a final decision fall or not fall. The network is retrained with the a public dataset FDD [17]. RetinaNet is used to detect the humans in the training videos of the dataset and then following features aspect ratio of bounding box, MHI, and head detector are extracted as described previously. For each frame of the training video, MHI is calculated. Both of the frame and MHI are used to feed first layer of MobileNet. Input RGB frames includes visual features and MHI represents motion features of humans in videos. The input frame and MHI are resized to be a vector of [224, 224,3] dimensions. The main idea of MobileNet models are based on depthwise separable convolution where each block consists of a 3x3 depthwise convolution that filters the input frames followed by a 1x1 pointwise convolution. MobileNet has 28 layers with counting depthwise and pointwise convolutions as separate layers followed by pooling and fully connected layer. Each layer was followed by batch normalization and ReLU activation. The input frame and MHI go through the first convolutional layer which it is a fully convolution layer then a 3 x 3 depthwise convolution followed by a 1 x 1 pointwise convolution is applied. Thereafter, these filtered values are combined to create the feature map. The output of the first layer is used to feed the next layer and the same previous operations of the first layer is applied to create a feature map but with small size and deep depth than the previous layer. The process continues until it reaches to the pooling layer which scales the feature map to smaller size and deeper depth to reach predefined size (e.g. 7×7) in order to avoid over-fitting. For classification the input to the dense layer is the output from the final Pooling Layer which is flattened and then fed into the dense layer. Finally, SoftMax is used in the output layer in order to transform the output values into the range [0,1] and gives the final decision for the action fall or not-fall. Fig.8 illustrates the architecture of MobileNet.

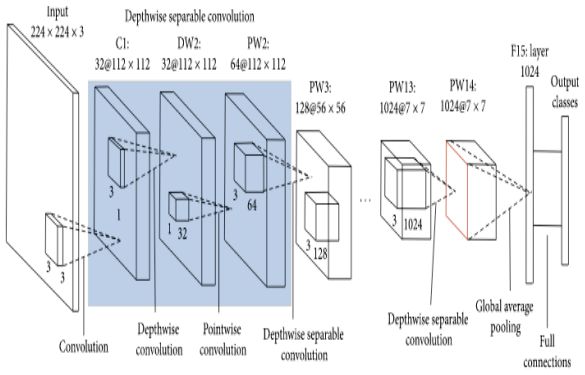

Fig. 8. Architecture of MobileNet.

IV. EXPERIMENTAL RESULT

Training and testing of the proposed framework are implemented using python version 3.0, TensorFlow, and Keras. Framework is trained online on Google Colab (colabrarory virtual machine) with free GPU and memory resources on cloud. The testing and the evaluation values of the proposed framework have resulted from epoch 200 of the trained model. The proposed framework is tested and evaluated using two datasets: UR Fall Detection (URFD) dataset and Fall Detection Dataset FDD. UR Fall Dataset (URFD) includes 30 videos of falls and 40 videos for daily activities life refer as no falls. FDD dataset includes 191 videos with frame rate of 25 frames/s and the resolution is 320x240 pixels [18]. The videos frames have three channels (i.e. R, G, and B) and are resized to 224x224x3. FDD dataset is recorded in different locations allowing to define several evaluation protocols (”Home”, ”Coffee room”, ”Office” and ”Lecture room”). FDD dataset was selected to train the proposed framework because it keeps the coordinates of the detected person on each frame and the coordinates of the fall action. Therefore, the training is conducted on a small portion of the dataset videos (”Coffee_room_01”, ”Coffee_room_02”, ”Home_01”, ”Home_02”). The datasets video sequences contain challenges like illumination change, occlusions, and textured background. There is only one actor in a video.

MobileNet trained for 200 epochs and the loss function is evaluated for each epoch during the learning process for both the training and the validation data. The proposed method applies early stopping at 30 epochs because loss does not improve. This is an ideal training strategy to avoid full training when it is not required. For loss function, the value of 1 is used for w1 class weight because increasing ’no fall’ category is not required. Fig.9 shows the loss obtained from the training and validation at each epoch. As shown in the figure, when the epoch increases, both the training and validation values are going close to 0 with the convergence of both of the training and validation values in output value which is considered as an accuracy indication of the proposed framework. The testing and the evaluation values of the proposed framework have resulted from epoch 200 of the trained model which takes around 72 minutes to train 34 epoch of the training FDD dataset. The experiment was conducted with the following Hyper-parameters shown in table (1).

Parameter	value
EPOCH	200
LEARNING_RATE	0.01
MOMENTUM	0.95
WEIGHT_DECAY	0.001
BATCH_SIZE	32

TABLE I. PARAMETER SETUP FOR THE TRAINING OF MOBILNET

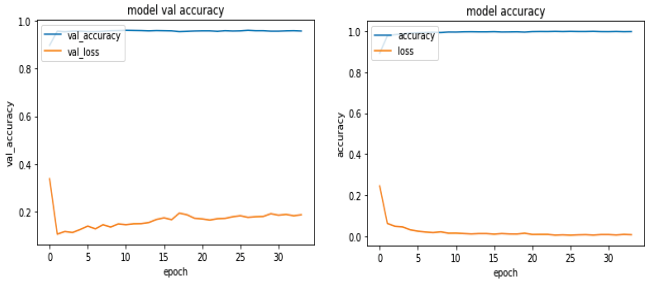

Fig. 9. The training and validation loss per epoch.

A. Qualitative Evaluation

To assess the proposed method efficiency, the experimental results of the proposed framework are compared with several fall detection methods qualitatively and quantitatively. The proposed framework is compared with other approaches which used hand-crafted features such as Yu al. approach [19], Charfi et al [20], and Rougier et al [21]. In addition, the proposed framework is compared to fall detection approaches based on CNN such as Adrian et al.[22] and Fakhruddin et al. [23]. Fig.10 shows visual results comparison in different frames and scenes among the proposed framework and methods described previously. The first row of Fig.10 is the input frames generated from the video scenes, second row is the ground-truth used for evaluation process, the next rows are the results of different comparable methods, and the last row shows the results of the proposed framework. As shown in Fig.10, The results of Yu et al [19], Charfi et al [20], and Rougier et al [21] methods which based on hand crafted feature extraction appear to be unstable and give poor results in case of occlusion problem and high miss-classifications between human and other objects in scene. Adrian et al.[22] approach is based on optical flow images which involves heavy computational complicity for pre-processing consecutive frames and fail to detect falls in case of illumination changes. Fakhruddin et al. approach suffer from the requirement of a large amount of data to detect a fall which consumes a lot of network bandwidth. The proposed framework provides accurate human fall decision and visually outperforms several comparable methods in different circumstances such as occlusion and illumination changes.

B. Quantitative Evaluation

To assess the efficiency of the proposed framework against comparable fall detection methods, accuracy, sensitivity, specificity and precision metrics are used. To calculate these metrics, the parameters true positive (TP) which represents the number of fall events detected as fall, false positive (FP) which represents the number of non-fall events detected as fall events (also known as false alarms), true negative (TN) which represents the number of non-fall events detected as non-fall events, and false negative (FN) which represents the number of

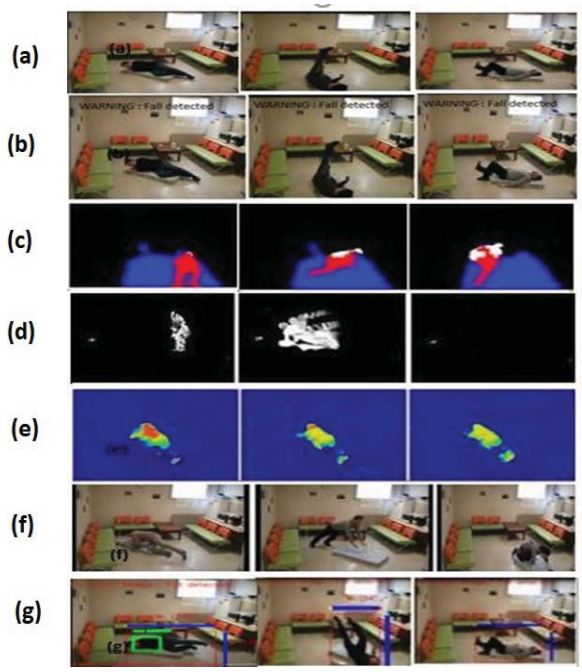

Fig. 10. Visual results comparison (a) Input frames, (b) Ground truth frames, (c) Detected outputs resulted from Yu et al[19], (d) Detected outputs resulted from Rougier et al [21], (e) Detected outputs resulted from Adrian et al.[[22], (f) Detected outputs resulted from Charfi et al [20], and(g) Detected outputs of the proposed method.

fall events detected as non-falls events. Accuracy, sensitivity, specificity and precision metrics metrics are calculated as follows:

$$Accuracy = (TP + TN)/(TP + FP + FP + FN) \quad (2)$$

$$Senstivity/recall = (TP)/(TP + FN) \quad (3)$$

$$Specificity = TN/(TN + FP) \quad (4)$$

$$Percision = TP/(TP + FP) \quad (5)$$

Fig.11 shows the confusion matrix obtained using the proposed framework based on FDD dataset. Although there appears a confusion between “Fall” and “not Fall” caused by falling-like activities in some testing videos, but most of the samples are correctly classified and the proposed framework achieves an overall accuracy of 98.8%.

Several comparisons with the the state-of the-art fall detection methods are conducted to ensure the robustness of the proposed framework based on UR and FDD datasets. The proposed framework outperforms the results of Yu et al [19], Charfi et al [20], and Rougier et al [21] methods as the proposed framework uses RetinNet to detect humans and avoid any misclassifion in stillness case as well as to avoid the overlap between background and detected humans. In addition, extraction of hand crafted features for well detected humans (due to the efficiency of RetinNet) to train MobileNet

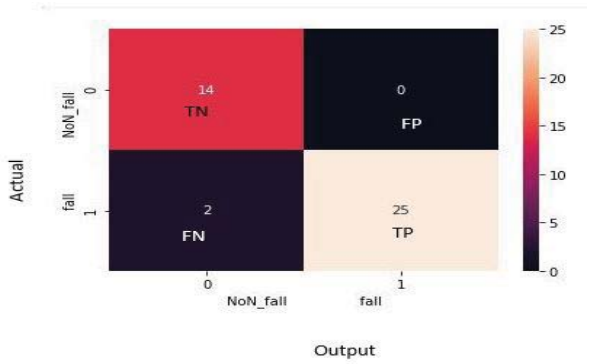

Fig. 11. Confusion matrix of the FDD dataset.

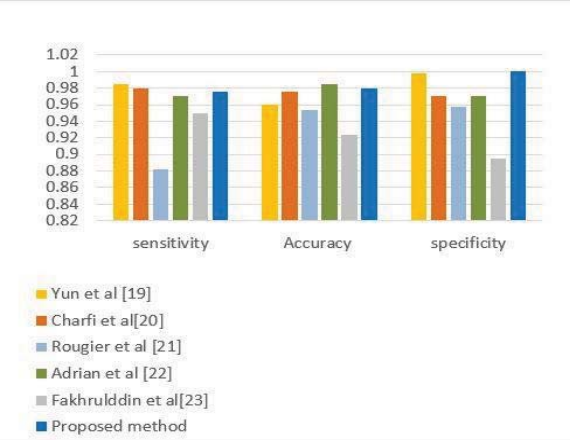

Fig. 12. Accuracy, sensitivity, and specificity Comparison among the proposed framework and state-of- art fall detection methods .

network contributes in achieving more accuracy. The proposed framework employs MobileNet as a classifier which is 30 times smaller and about 10 times faster compared with VGG which used in the approach presented in [22]. Depthwise convolution in MobileNet proves high accuracy with low-latency and low-power with model performance of 0.04 fps. The proposed framework achieve 98.5% accuracy on test data where the framework was trained on ”Coffee_room_01”, ”Coffee_room_02”, ”Home_01”, ”Home_02 ” as a part of used data. The proposed framework accuracy is the highest in comparison to the work presented in [23] which achieved 92.3%. Furthermore, the proposed framework achieved 98% on analyzing of the sensitivity for lying pose which is extremely desirable in fall detection where an after-fall pose is considered to be lying. Similarly, the proposed framework shows high precision regardless of the human position, illumination changes, or facing any camera angle side. Fig.12 illustrates comparison among the proposed framework and other state-of the-art fall detection methods.

V. Conclusion

In this paper, a proposed framework was presented for human fall detection based on the combination of convolutional neural networks and handcrafted features. The proposed method is based on analyzing the human appearance, motion,

and the shape variations in videos. The proposed fall detection is performed in three phases: human detection, feature extraction, and action classification to fall or not-fall. RetinaNet is a one-stage object detection CNN that used for accurately determining the position of humans in videos. The feature extraction is the second phase of the proposed framework in which features that represent the detected human motion and body shape (MHI, aspect ratio, and head detection) are extracted. These features form a feature map that is used to feed MobileNet model. The third phase of proposed method used a modified version of MobileNets to classify human motion according to the features extracted in the second phase into "fall" or "not fall". Experimental results showed the efficiency of the proposed framework to detect human falls in different scenarios and situation achieving up an accuracy up to 98%. In the future work, the feature extraction using convolutional neural networks are intended to be used instead of handcrafted features accompanied with an analysis of the computational time to the proposed framework with handcrafted features and CNN feature extraction.

REFERENCES

[1] The prevention of falls in later life. A report of the Kellogg International Work Group on the Prevention of Falls by the Elderly. Dan Med Bull. 1987 Apr;34 Suppl 4:1-24. PMID: 3595217.

[2] Muhammad Mubashir, Ling Shao, Luke Seed,"A survey on fall detection: Principles and approaches,"Neurocomputing, Volume 100,2013,Pages 144-152,ISSN 0925-2312,.

[3] T. Tri, H. Truong, and T. Khanh, "Automatic Fall Detection using Smartphone Acceleration Sensor," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 12, pp. 123–129, 2016.

[4] W. Lie, A. T. Le and G. Lin, "Human fall-down event detection based on 2D skeletons and deep learning approach," 2018 International Workshop on Advanced Image Technology (IWAIT), Chiang Mai, 2018, pp. 1-4, doi: 10.1109/IWAIT.2018.8369778.

[5] Khan, M. and H. A. Habib. "Video Analytic for Fall Detection from Shape Features and Motion Gradients." (2009).

[6] Yu, M. et al. "A Posture Recognition-Based Fall Detection System for Monitoring an Elderly Person in a Smart Home Environment." IEEE Transactions on Information Technology in Biomedicine 16 (2012): 1274-1286.
bibitem7 X. Yu, "Approaches and Principles of Fall Detection," Healthcom, vol. 10, no. 1, pp. 42–47, 2008.

[7] N. Zerrouki and A. Houacine, "Combined curvelets and hidden Markov models for human fall detection," Multimedia Tools and Applications, pp. 1–20, 2017.

[8] A. Doulamis and N. Doulamis, "Adaptive deep learning for a vision-based fall detection," in Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference. ACM, 2018, pp. 558–565.

[9] A. Núñez-Marcos, G. Azkune, and I. Arganda-Carreras, "Vision-based fall detection with convolutional neural networks," Wirel. Commun. Mob. Comput., vol. 2017, 2017.

[10] Ghiasi G, Lin T-Y, Le QV (2019) NAS-FPN: learning scalable feature pyramid architecture for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7036–7045.

[11] A. G. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile VisionApplications," 2017.

[12] Ge C, Gu I Y and Yang J 2017 Human fall detection using segment-level cnn features and sparse dictionary learning IEEE 27th Int. Workshop on Machine Learning for Signal Processing (MLSP) pp 1-6.

[13] Chua, JL., Chang, Y.C., and Lim, W.K. A simple vision-based fall detection technique for indoor video surveillance. SIViP 9, 623–633 (2015).

[14] K. Sehairi, F. Chouireb, and J. Meunier, "Elderly fall detection system based on multiple shape features and motion analysis," 2018 International Conference on Intelligent Systems and Computer Vision (ISCV), Fez, 2018, pp. 1-8, doi: 10.1109/ISACV.2018.8354084.

[15] Geismann, P. and G. Schneider. "A two-staged approach to vision-based pedestrian recognition using Haar and HOG features." 2008 IEEE Intelligent Vehicles Symposium (2008): 554 559.

[16] Y. Li, H. Huang, Q. Xie, L. Yao, and Q. Chen, "Research on a surface defect detection algorithm based on MobileNet-SSD," Appl. Sci., vol. 8, no. 9, 2018.

[17] fall detection Dataset .Available online " http://le2i.cnrs.fr/Fall-detection-Dataset " Antoine Trapet - 27 February 2013.

[18] M. Yu, A. Rhuma, S. M. Naqvi, L. Wang and J. Chambers, "A Posture Recognition-Based Fall Detection System for Monitoring an Elderly Person in a Smart Home Environment," in IEEE Transactions on Information Technology in Biomedicine, vol. 16, no. 6, pp. 1274-1286, Nov. 2012, doi: 10.1109/TITB.2012.2214786.

[19] I. Charfi, J. Miteran, J. Dubois, M. Atri and R. Tourki, "Definition and Performance Evaluation of a Robust SVM Based Fall Detection Solution," 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems, Naples, 2012, pp. 218-224, doi: 10.1109/SITIS.2012.155.

[20] C. Rougier, J. Meunier, A. St-Arnaud and J. Rousseau, "Fall Detection from Human Shape and Motion History Using Video Surveillance," 21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07), Niagara Falls, Ont., 2007, pp. 875-880, doi: 10.1109/AINAW.2007.181. 875880.

[21] Adrian Núñez-Marcos, G. Azkune, and I. Arganda-Carreras, "Vision-based fall detection with convolutional neural networks," Wirel. Commun. Mob. Comput., vol. 2017, 2017.

[22] Fakhrulddin A., Fei X., and Li Hanchao. (2017). Convolutional neural networks (CNN) based human fall detection on Body Sensor Networks (BSN) sensor data. 1461-1465. 10.1109/ICSAI.2017.8248516.