Int J Syst Assur Eng Manag (March 2023) 14(Suppl. 1):S79–S86
https://doi.org/10.1007/s13198-022-01822-y

ORIGINAL ARTICLE

Recognizing elderly peoples by analyzing their walking pattern using body posture skeleton

Dushyant Kumar Singh[1]

Received: 3 February 2022 / Revised: 4 July 2022 / Accepted: 26 November 2022 / Published online: 1 December 2022

Abstract The increasing age of the population has become a significant concern internationally. During the COVID-19 pandemic situation, it has been seen that the most sensitive and affected class of the population is the class of Elder's. It is therefore necessary to track the movement and behavior of the old persons. This kind of monitoring could help them in providing assistance in their needy time. Our objective is to develop an approach to classify elderly people using skeleton data for their assistance. OpenPose algorithm is used here to detect human skeletons (joint positions) from the video sequences. OpenPose algorithm with a sliding window of size 'N' is used to achieve a real-time posture recognition framework. Posture features from each extracted skeleton are then used to build a classifier for recognizing elderly people. We also introduce here a new dataset that includes old person walk and young person walk video's. The experimental outcomes reveal that the proposed method has achieved up to 98.45% training accuracy and 96.16% testing accuracy for deep feed-forward neural network (FFNN) classifier. This asserts the effectiveness of the approach.

Keywords Feed forward neural network (FFNN) · OpenPose · Human detection · Posture recognition · Skeleton

1 Introduction

According to the 2011 Population Survey (Velayutham et al. 2016), approximately 104 million people were aged 60 years or older. Another report (Singh and Kaur 2018) reveals that the elderly population in India is supposed to increase to 173 million by 2026. In 2010, Central Intelligence Agency World Factbook (Agency et al. 2010) recorded the elderly dependency ratio of 9.8 of the elderly people of age 65 or above per 100 working people between ages 15–64. Increased aging dependency ratio puts additional pressure on caregivers. Since elderly individuals are more prone to diseases, their health care therefore is a serious concern (Lubeek et al. 2017).

The work under this manuscript focuses on enhancing the quality of life of elderly people and preventing them from any unwanted accidents related to their health status. This can be done through real-time monitoring of elder ones either physically by ourselves or through some digital technology in case of our absence. The two prevalent tools of digital technology that can be used for monitoring any movement of elderly people are "Sensors based assistive device(s)" and/or "Computer vision based techniques". The first way is to deploy wearable devices in different body parts of the person to detect lively activities of daily routine. These devices are usually placed behind the ear lobe, under the axilla, around the wrist, or at the waist. It is a tiresome job to wear them all the time. The problem mostly arises from improper use of such detectors, as people mostly forget to wear them. This behavior of humans limits the performance of these detectors. On the other hand, the computer vision system tries to excerpt some noticeable postures and features from the videos of the elderly people's movements and analyze these for activity-like patterns. The recognized activity can then help assessing the health condition of elder ones.

The computer vision system attracts a great deal of interest, specifically in assistive technology for elderly people. Computer vision systems perform several tasks like

✉ Dushyant Kumar Singh
dushyant@mnnit.ac.in

[1] MNNIT Allahabad, Prayagraj, India

detecting human presence and recognizing elderly people based on their movement. This could ensure the safety and comfort of all the elderly people living on their own. The primary health problem of elderly people is that they are prone to fall, which leads to very long-term injuries, fear and even death in some cases. Falling incidents lead to fractures and psychological consequences that lessen their independence. According to the study (Visutsak and Daoudi 2017), 28–34% of older persons fall at least once a year, of which 40–60% of those falls resulting in injury. Therefore, this paper proposed a practical assistive-technology based surveillance system to identify young and old people's in real-time video sequences. One of the popular pose estimation techniques, namely the OpenPose algorithm (Cao et al. 1812), is used here to derive skeletons in terms of posture joints. Since the body movements of younger and older people are different, they can be identified as younger or older based on joint movements.

The list of the contributions presented in this article is as follows:

(a) We present an activity recognition framework that analyzes 2D skeletal data and classifies related actions.
(b) We present skeleton pre-processing and feature extraction methods to extract relevant features from a sequence of skeletal data.
(c) We perform a variety of experiments on a synthesized video dataset to access the walking patterns of elderly people.

Computer vision based proposed assistance mechanism is beneficial for the caregivers to take care of older people at homes or hospitals. The remaining of the article is organized as follows. After the brief literature discussion in Sect. 2, Sect. 3 describes the proposed methodology for recognizing older people based on their movements. In addition, this section also includes a brief discussion of the various techniques involved as part of the proposed methodology. The experiments and discussion are presented in Sect. 4. Finally, the last section involves the conclusion of the paper.

2 Related works

Nowadays, researchers are widely adopting sensor-based and vision-based approaches to recognize human acts in real-time scenarios. The sensor-based approach uses wearable sensors like accelerometers, gyroscopes, etc., to track an individual's activity. In contrast, vision-based approaches mostly use a Convolutional neural network (CNN) to classify human activities in real-time video sequences. Convolution neural network (CNN) (Ansari and Singh 2021; Ojha et al. 2017) is an influential innovation in computer vision. A CNN is a deep learning approach that automatically learns spatial hierarchies of features through back propagation by using input, Convolutional, pooling, fully connected, and output layers. Some of the researches based on sensors and vision are discussed as follows:

Pienaar and Malekian (2019) suggested human activity recognition (HAR) system to track the basic activities of a person by analyzing raw sensor data. Here, an open-source dataset introduced by Wireless Sensor Data Mining Lab is used that consists of six activity levels. The outcomes show that the model has achieved an accuracy of more than 94. Chernbumroong et al. (2013) suggested a HAR system to detect activities of daily living by analyzing wrist-worn sensor data. The data produced by sensors is first pre-processed and then passed to the feature extraction module to extract relevant features. Next, the extracted features are used to build a classifier to categorize the involved activities. This method shows proficient outcomes in terms of accuracy up to 94%. Putra and Yulita (2019) proposed HAR model based on the bed-wake gesture. Here, a multilayer perceptron network was used to predict activity based on sensor readings. This work has achieved proficient results in terms of accuracy of up to 90.17% for MLP and 84.46% for Naïve Bayesian.

Ma et al. (2019) employed two-stream deep ConvNets to build an expert HAR system system using Inception-style temporal Convolutional Neural Network and a Recurrent Neural Network. Both networks are used to extract spatiotemporal information and exploit Spatio-temporal dynamics to enhance the whole system's performance. This method has provided excellent accuracy, up to 94.1% for on UCF101 dataset and 69.0% for HMDB51 dataset. Ji et al. (2012) developed a novel 3D Convolution Neural Network for recognizing human activities in surveillance videos. The deep 3D-CNN model evaluates appearance and motion features from each video frame. This method has attained superior performance with an average accuracy of 90.2%. Zhang and Tian (2012) study various Spatio-temporal features-based descriptors for activity recognition. They found that the probabilistic graphical models are good enough to recognize activity patterns over time than SVMs. Li et al. (2018) used deep neural networks to recognize various actions based on modeling human body posture. The method integrates RGB and optical flow streams with 2D posture features to perform human activity classification.

The literature, as discussed in sensor-based HAR systems, requires wearable sensors such as accelerometers, glucometers, proximity sensors, etc., to recognize human actions. However, the range of detecting human actions is limited in sensors-based HAR systems. Other side, vision-based HAR systems can identify a wide range of human acts using camera-based surveillance. They use complex convolutional neural architectures to learn temporal relations for human acts. However, improvements are still being required to enhance the performance of existing HAR systems. Therefore, this work proposed a cost-effective solution to differentiate human actions by analyzing human 2D body joints. The details related to the proposed system are presented in Sect. 3.

3 Proposed methodology

As mentioned earlier in the introduction section, this manuscript proposes an assistive technology based surveillance system to classify young and old persons in real-time scenarios. The overall workflow of the proposed system is presented in Fig. 1. The system takes video stream as input through the camera, examines each frame, and categorizes younger and older people by analyzing their walking styles. So that caregivers can be more attentive in the case of older people. The OpenPose algorithm is used to generate the human skeleton (pair of joints locations) in each frame. The OpenPose algorithm inputs an RGB frame of size “$w \times h$” and provides the joint locations to form a skeleton for each individual within an image. After getting the skeleton joints, the skeleton data is aggregated for the first N frames using a sliding window of size N. Here, N skeletons are first pre-processed and then passed to a feature extraction module that extracts relevant features from them. Further, the extracted features are used to build a classifier for categorizing young and old walks. To achieve a real-time recognition framework, the window slides frame by frame along the video’s time dimension and outputs a label for each video frame.

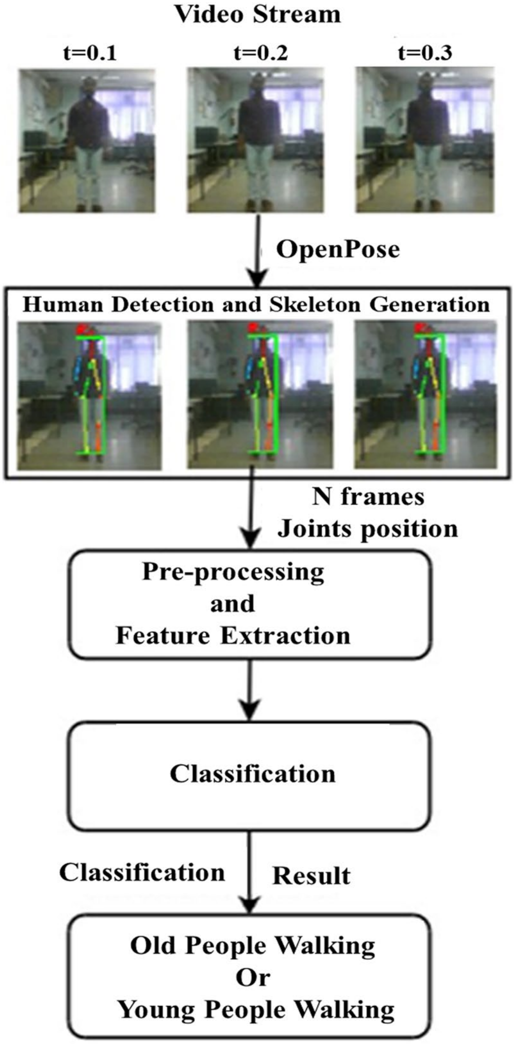

Video Stream
t=0.1
t=0.2
t=0.3
OpenPose
Human Detection and Skeleton Generation
N frames
Joints position
Pre-processing and Feature Extraction
Classification
Classification
Result
Old People Walking Or Young People Walking

Fig. 1 Workflow of the proposed system

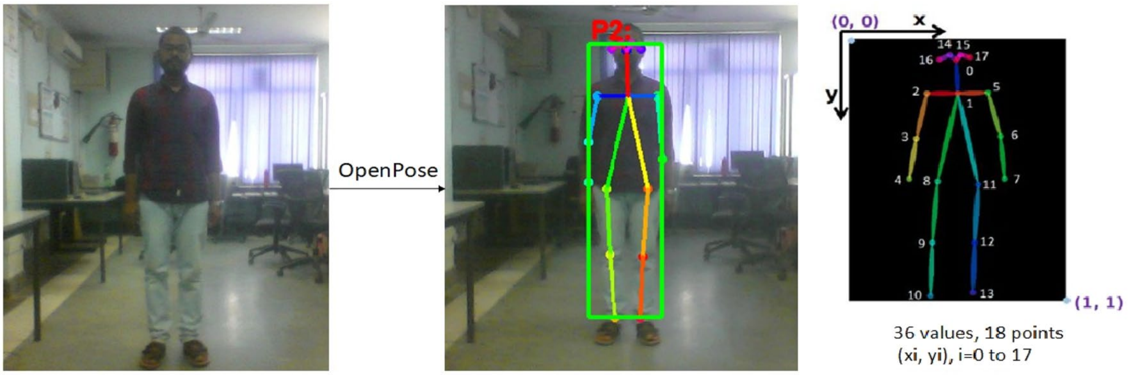
3.1 Human detection and skeleton generation

Human detection and skeleton generation are the primary tasks for identifying old and young walks. The work presented here uses the OpenPose algorithm (Cao et al. 1812) for human detection and skeleton generation from an image. It can jointly detect the human body and involve key points to generate skeleton. The OpenPose provides two Heat Maps, one for evaluating joint positions, i.e. Confidence Map (S), and the other for associating the joints, i.e. Part Affinity Field (PAF) Map (L) in a human skeleton. The OpenPose algorithm takes an image as input and spots skeletons for the person found in that image. An extracted skeleton involves 18 joints, including head, neck, arms, and legs, as shown in Fig. 2a. Each joint position is represented using spatial pair of coordinates, i.e. (x, y). Therefore, each skeleton is represented using 18 pairs of coordinates (a total of 36 values), as shown in Fig. 2b.

3.2 Pre-processing for features extraction

After extracting the raw skeleton, the pre-processing stage suppresses the unwanted distortion from the skeleton data. The pre-processing stage helps to enhance the characteristics of skeleton data, which helps in more accurate classification at later stage. The pre-processing includes four steps summarized as follows:

- *Considering all head’s joints:* Along with the body and limb configurations, the head position can help a lot for the classification. Therefore, the five joints on the head are added manually to make the features more meaningful.
- *Coordinate Scaling:* The x and y coordinates for representing joint position do not follow the same scale. Therefore, these points need to be normalized in the same unit to deal with different width and height ratios.
- *Discard frames that do not have neck and thighs:* If OpenPose does not recognize a human skeleton or if the



(a) (b)

Fig. 2 **a** Skeleton representation using OpenPose, **b** Representation of skeleton in spatial domain

identified skeleton does not have a neck or thighbone inside the frame, the frame is considered invalid and dropped. The sliding window slides to the successive frames.

- *Fill the missing joints:* OpenPose may fail to recognize a full human skeleton in an occluded environment, which results in blanks at joint positions. To keep a fixed-size feature vector for the classification purpose, these joints must be filled with certain values. Here, the position of the missing joint is determined by its relative position to the neck in the preceding frame.

3.3 Features extraction

After pre-processing, the joint positions are completed and ready for use in the feature extraction process. Therefore, we used the sliding window of size N with N = 5 to extract relevant features from extracted joint positions that help to identify the action types. A better presentation with skeleton from five consecutive frames is illustrated in Fig. 3.

The salient features are constructed using normalized joint positions by calculating the moving velocity of the joints and the angle of each joint for N window size. Further, a feature vector is created by concatenating these features, and then the extracted vectors are fed into a deep FFNN Classifier for training. The algorithm for finding more salient features from the raw skeleton data is discussed as follows:

Algorithm 1. Finding the more salient features from the raw skeleton data

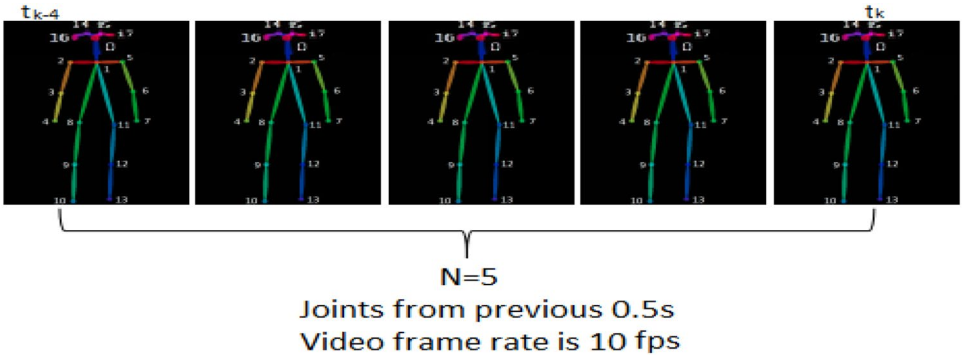
Result: A feature vector appended with the following feature.

1. Link all the joints for N frames, Where N = 5 is Sliding Window.
2. Average the height (H) of the skeleton by considering previous N frames. This height is equivalent to the distance between the neck and the thigh. It is used to normalize all of the characteristics listed below.
3. Normalize the joint positions (N), as presented in equation 1.

   (1) N = [Joint Positions - mean (Joint Positions)]/Height (H)
4. Velocity (V) between joints is calculated as the next joint position minus the previous joint position in the normalized coordinate, as presented in equation 2.

   (2) V = N[tk] - N [tk-1]
5. Compute the angle of each joint from the joint's positions.

Fig. 3 Raw joint positions



3.4 Deep feed-forward neural network

A deep feed-forward neural network (FFNN) is a deep neural network comprised of two or more layers of neurons. Feed-forward Network consists of an input layer and an output layer. The input layer is responsible for receiving the signal, while the predictions about the input are made in the output layer. There are an uncertain number of hidden layers between input and output layers in which the actual computation has to be performed. Only one hidden layer in FFNNs is proficient in approximating any continuous function. FFNN learns to simulate the correlation between inputs and outputs by training on a collection of input–output pairings. The model parameters, or weights and biases, are adjusted throughout training to reduce the error. Backpropagation is utilized to adjust the weights and biases relative to the error, and root mean squared error is used for measurements. FFNN updates the partial derivatives of the error function for many weights and biases using back propagation and the chain rule of calculus. Figure 4 shows FFNN architecture with hidden layers.

In this work, the deep FFNN model has been used to detect and recognize elderly people that contain one input layer, three hidden layers, and an output layer. Three times dropout has been used to prevent the model from overfitting. The input layer has 314 nodes which are features extracted from the raw skeleton data. The rectified linear unit (ReLU) is used here as an activation function to deal with non-linear input data. There are 100 nodes in each hidden layer and 2 nodes in the output layer. The output layer has 2 nodes for 2 different classes. The sigmoid function has been used as an activation function that is used to calculate the probabilistic value of each class with a learning rate of 0.0001. The highest probabilistic class will be considered as an output to the corresponding input image.

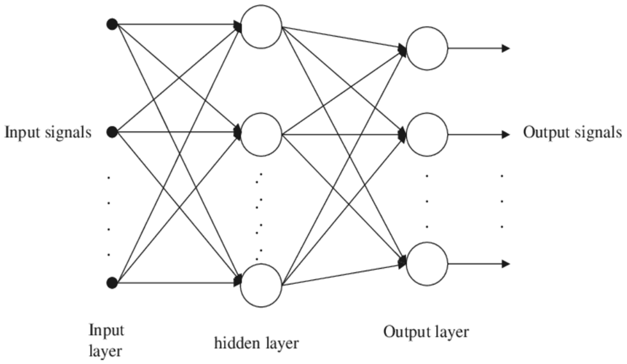


Fig. 4 FFNN architecture with hidden layers

4 Experimentation

Google Colab platform has been used to perform a wide range of experiments. PIL (python imaging library) and OpenCV have been employed to open, save, and manipulate images. Keras library is used for classification purposes by incorporating SVM with linear/RBF kernel and Deep Neural Network. The Matplotlib library is used here to visualize model accuracy and loss curves. Scikit-learn is employed to produce the confusion matrix, and TensorFlow is used as a data flow.

4.1 Dataset

Dataset is synthesized by ourselves in an indoor and outdoor environment using a 16 MP Mobile camera. The dataset contains two types of people walk such as elderly people walk and younger people walk. Each class consists of a variable amount of videos, ranging from 30 s to 2 min in length. Videos are captured at the resolution of 640 × 480. For machine understanding, the created dataset requires proper formatting and labeling for training the model. YAML (Yet Another Markup Language), a comprehensible information serialization language is used for this purpose. The images that are going to be used for training and their label are configured in a text file containing the information like class name, starting and ending index of the video corresponding to that class. The distribution of our dataset is presented in Table 1.

Figure 5 shows some instances/samples of video sequences representing old people walking and young people walking. The clips are shot in different scenarios like indoor, outdoor, low lighting, etc.

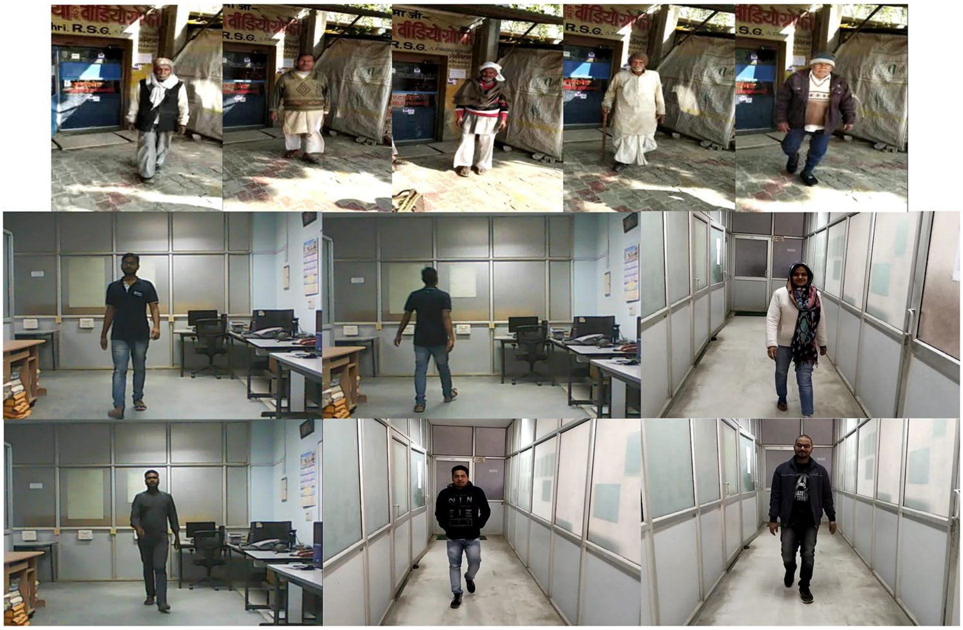
4.2 Training

The entire dataset is divided as 70% for training and 30% for testing. After pre-processing and features extraction from the raw skeleton data, the next phase forms a model to classify the data. The classification has been done using different classifiers, including Neural Network (FFNN Classifier), Support Vector Machine (SVM), and SVM with kernel method. Setting up the hidden layer and balancing the learning rate (Putra and Yulita 2019) are worked out for efficient modeling. To obtain the best outcome for each parameter, the experiment was repeated several times to find the best amount of hidden layers and the best learning rate. For updating the parameters, the loss function partial derivate

Table 1 Dataset's frame distribution

Actions	Old people walk	Young people walk
No. of frames	4526	5246

Fig. 5 Snapshots of two types of classes in the training data: elderly people and young people walk



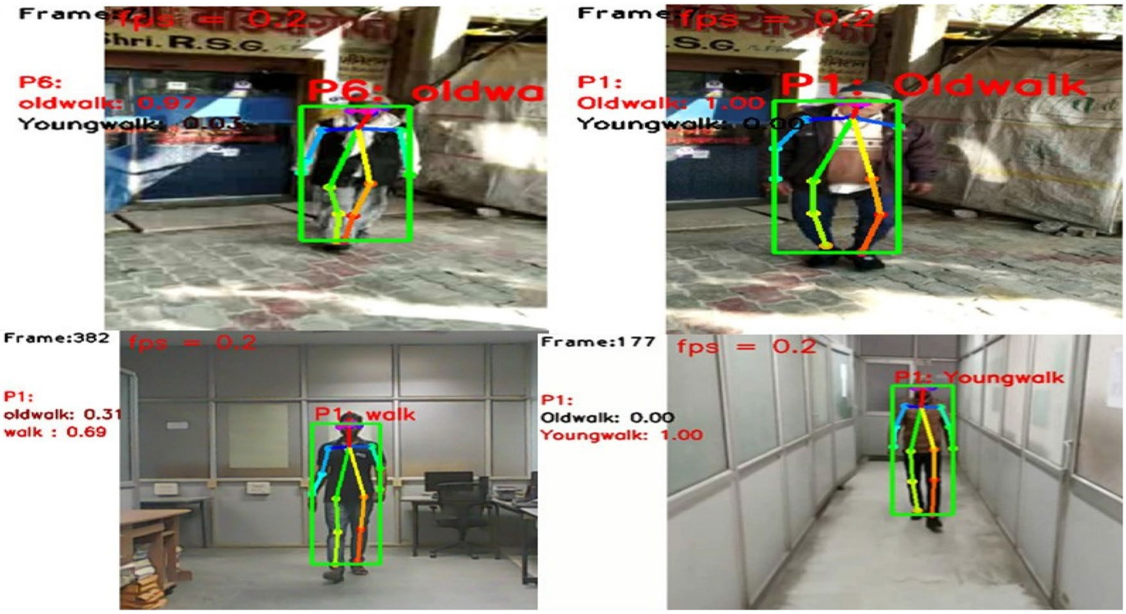


Fig. 6 Sequences outcomes

w.r.t the model parameters is computed. Validation is done along with training. Figure 6 illustrates the performance of a posture recognition system on a synthesized dataset where people perform the walking action. The trained model is fine enough for detecting the old walk or young walk.

4.3 Testing

The model is tested for 2930 test images containing two classes, each having 1357 and 1573 images to represent old people walk and young people walk, respectively. The test images are passed to our method proposed for prediction. The model first processes each test image through the OpenPose to detect the human skeleton in that image. The skeleton data is passed to pre-processing module, feature

Table 2 Recognition accuracy for N = 5

Method	Classifier	Accuracy train (%)	Accuracy test (%)
SVM	Linear	88.8	89.1
SVM	Kernel	91.6	92.1
FFNN classifier	100 × 100 × 100	98.45	96.16

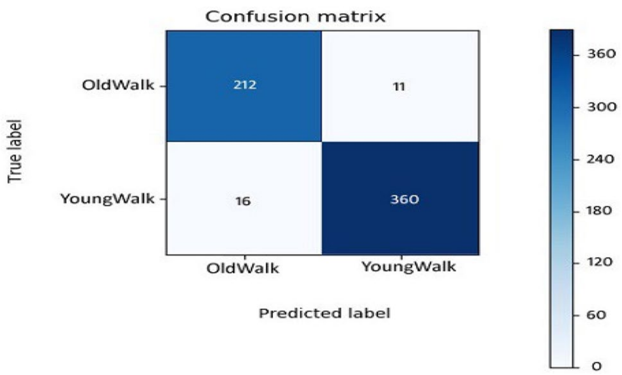


Fig. 7 Confusion matrix

extraction module, and classification module. The model weights are adjusted during the training phase, and the window slides frame by frame along the video's time dimension in which the prediction has been made. The proposed recognition system's performance for elderly people has been tested on our dataset. Testing videos are also similar to training videos.

4.4 Result

This section shows the experimentation outcomes of the proposed model. Table 2 shows the recognition accuracy of the proposed method over different classifiers. The outcomes show that the proposed method trained over the FFNN classifier is quite higher than both variants of the SVM classifier.

The Confusion matrix of the Deep Neural Network model is given in Fig. 7. Diagonal values of the matrix represent correctly classified testing outcomes. Non-diagonal values represent miss-classified outcomes. Miss-classified means predicted value and actual value are not matching.

Different performance measures (Singh 2015, 2017; Reddy and Geetha 2020) like Precision, Recall, F-Measures, and Support are used to evaluate the performance of this proposed model, illustrated in Table 3.

The performance of the proposed method is compared with other existing methods in Table 4. The result drawn in Chernbumroong et al. (2013) works on analyzing the sensors' data for assisted living and provides an accuracy of 90.23%. In Reddy and Geetha (2020), activities are modeled using video-based classification that offers up to 92.16% accuracy. However, the proposal in Li et al. (2018) classifies activities by modeling body posture using a deep neural network and stands good compared to others with an accuracy of 93.61%. The last and our proposed method achieves around 96% of accuracy for almost the same activities and behavioral modeling.

Table 3 Evaluated performance measures

Metrics name	Old people walk	Young people walk
Precision	0.94	0.98
Recall	0.96	0.96
F-measure	0.95	0.97
Support	223	376

Table 4 Comparison with existing methods

Methods	Accuracy (%)
Chernbumroong et al. (2013)	90.22
Reddy and Geetha (2020)	92.16
Li et al. (2018)	93.61
Proposed method	96.16

5 Conclusion

These manuscripts proposed a system to spot human presence and recognize whether the particular person is an older adult by analyzing the human walking style. This system constructs features from the pre-processed skeleton data constructed over video sequences. The developed method aggregates the skeleton data of a 0.5 s window for feature extraction. The model used raw features from five consecutive frames to improve the models' performance. We evaluated the developed model over our synthesized dataset. This paper considered a deep neural network model to recognize elderly people in the video. There are also two other variants of the SVM classifier, such as SVM with linear kernel and RBF kernel, which achieves good accuracy in elderly people's recognition. The result shows that the deep neural network model outperforms Linear SVM and RBF-SVM on our dataset, demonstrating good results in real-world environments. In the future, the system can be used to deploy in different applications like smart homes, theft detection, augmentation reality and many more. Additionally, more adaptive techniques like advanced CNN architecture can be used to upsurge the performance of the proposed system.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Agency CI (2009) The CIA world factbook 2010. Skyhorse Publishing Inc.

Ansari M, Singh DK (2021) Human detection techniques for real time surveillance: a comprehensive survey. Multimed Tools Appl 80(6):8759–8808

Cao Z, Hidalgo G, Simon T, Wei S-E, Sheikh Y (2018) Openpose: real-time multi-person 2d pose estimation using part affinity fields. arXiv preprint arXiv:1812.08008

Chernbumroong S, Cang S, Atkins A, Yu H (2013) Elderly activities recognition and classification for applications in assisted living. Expert Syst Appl 40(5):1662–1674

Ji S, Xu W, Yang M, Yu K (2012) 3d convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231

Li C, Tong R, Tang M (2018) Modelling human body pose for action recognition using deep neural networks. Arab J Sci Eng 43(12):7777

Lubeek SF, van Vugt LJ, Aben KK, van de Kerkhof PC, Gerritsen M-JP (2017) The epidemiology and clinicopathological features of basal cell carcinoma in patients 80 years and older: a systematic review. JAMA Dermatol 153(1):71–78

Ma C-Y, Chen M-H, Kira Z, AlRegib G (2019) TS-LSTM and temporal-inception: exploiting spatiotemporal dynamics for activity recognition. Signal Process Image Commun 71:76–87

Ojha U, Adhikari U, Singh DK (2017) Image annotation using deep learning: a review. In: 2017 international conference on intelligent computing and control (I2C2). IEEE

Pienaar SW, Malekian R (2019) Human activity recognition using LSTM-RNN deep neural network architecture. In: 2019 IEEE 2nd wireless Africa conference (WAC). IEEE, pp 1–5, IEEE

Putra D, Yulita I (2019) Multilayer perceptron for activity recognition using a batteryless wearable sensor. In: IOP conference series: earth and environmental science, vol 248. IOP Publishing, p 012039

Reddy GP, Geetha MK (2020) Video based fall detection using deep convolutional neural network. Eur J Mol Clin Med 7(2):5542–5551

Singh DK (2015) Recognizing hand gestures for human computer interaction. In: 2015 international conference on communications and signal processing (ICCSP). IEEE

Singh DK (2017) Gaussian elliptical fitting based skin color modeling for human detection. In: 2017 IEEE 8th control and system graduate research colloquium (ICSGRC). IEEE

Singh NP, Kaur G (2018) Urinary tract infection in elderly: to treat or not to treat? J Assoc Physicians India 66:11

Velayutham B, Kangusamy B, Joshua V, Mehendale S (2016) The prevalence of disability in elderly in India—analysis of 2011 census data. Disabil Health J 9(4):584–592

Visutsak P, Daoudi M (2017) The smart home for the elderly: perceptions, technologies and psychological accessibilities: the requirements analysis for the elderly in Thailand. In: 2017 XXVI international conference on information, communication and automation technologies (ICAT). IEEE, pp 1–6

Zhang C, Tian Y (2012) RGB-D camera-based daily living activity recognition. J Comput Vis Image Process 2(4):12

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.