Smart healthcare system-a brain-like computing approach for analyzing the performance of detectron2 and PoseNet models for anomalous action detection in aged people with movement impairments

R. Divya^{1,2} · J. Dinesh Peter¹

Received: 2 December 2020 / Accepted: 1 March 2021 / Published online: 10 March 2021
© The Author(s) 2021

Abstract

In this era of artificial intelligence, a wide variety of techniques are available in healthcare industry especially to study about various changes happening in the human body. Intelligent assistance using brain-like framework helps to understand and analyze various types of complex data by utilizing most recent innovations such as deep learning and computer vision. Activities are complex practices, including continuous actions as well as interleaved actions that could be processed with fully interconnected neuron-like processing machine in a way the human brain works. Human postures have the ability to express different body movements in different environments. An optimal method is required to identify and analyze different kinds of postures so that the recognition rate has to be increased. The system should handle ambiguous circumstances that include diverse body movements, multiple views and changes in the environments. The objective of this research is to apply real-time pose estimation models for object detection and abnormal activity recognition with vision-based complex key point analysis. Object detection based on bounding box with a mask is successfully implemented with detectron2 deep learning model. Using PoseNet model, normal and abnormal activities are successfully distinguished, and the performance is evaluated. The proposed system implemented a state of the art computing model for the development of public healthcare industry. The experimental results show that the models have high levels of accuracy for detecting sudden changes in movements under varying environments.

Keywords Activity modeling · Brain computation · Intelligent health caring · Fall detection · Deep learning · Keypoint detection

Introduction

According to World Health Organization, falls are considered as the second major cause of accidental or inadvertent injury deaths throughout the world. The number of falls results in death is more in the case of elderly people with more than 65 years old. Also millions of falls among them are severe enough which require serious medical considera-

tion [1]. These types of deaths are more in countries with a turnover middle or lower level. Avoidance techniques ought to be emphasized in education, training, making more secure conditions, prioritized researches in fall-related events as well as arranging powerful measures to lessen the risk of falls. Because of advances in clinical researches and innovations, the old populace has kept on developing. Services to the elderly people become a significant issue that ought to be dealt with cautiously. Intelligent assistance makes it very much helpful in healthcare fields. Healthcare informatics is an integrative area which incorporates the investigation of computer-based developments in medical assistance conveyance and management. An intelligent framework helps to understand and analyze various types of complex data in a simpler manner utilizing the most recent innovations such as deep learning and computer vision. An intelligent care tak-

✉ R. Divya
divya.r.sree@gmail.com

¹ Department of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamilnadu, India

² Department of Computer Science and Engineering, Sahrdya College of Engineering and Technology, Thrissur, Kerala, India

ing system provides continuous monitoring of people without much human intervention. The system needs to track different types of body movements in real-time irrespective of the environment. The main intention of this study is to adopt a futuristic method for human posture recognition and classification that uses video frames to develop an efficient posture recognition system in various environments. Different types of human posture detection and gait analysis techniques are available. Traditional methods use an ordinary camera to capture 2D images. Several geometric transformations have to be applied to the captured images. Problems associated with these methods are movements in the human body, changes in silhouette information, camera issues, image distortions due to dynamic background, and complexity in converting a 2D image to a 3D human body representation.

Activities are complex practices including continuous actions as well as interleaved actions. Human postures have the ability to express different body movements in different environments. An optimal method is required to identify and analyze different kinds of postures so that the recognition rate has to be increased. Low-cost motion sensing as well as depth-sensing devices such as Microsoft Kinect is currently being available for this purpose which results in the posture recognition process faster [2]. Researches in [3] obtained human detection based on 2D images captured using Kinect camera which provides information about human head contour. The nearest neighbour algorithm is used to get a meaningful array of data. Two-dimensional shape comparing algorithms are used for identifying the presence of the human body in the image. The standard height of the head concerning the depth is calculated. The presence of the human body in the image is located by using a 2D distance matching algorithm. Resultant depth information is processed to extract the edge points using edge detection algorithms. Regression analysis is done to compare the depth and height of the head. Use a 3D head model to extract generalized features of the head from different viewpoints. The general objective of fall forecasting is to forestall risky falls if possible.

Human activity modeling

Activity recognition is focused to identify and classify the nature of different human postures and activities from continuous stream of videos in real time. Brain-like computing systems are evolved to solve the problems that are highly pertinent to machine learning and robotics for movement detection and action recognition. Because of the progress in development of self-activated expounding of human postures, especially in the field of automated surveillance, and care taking systems, several researches are happening to enhance the existing recognition systems. The system should handle ambiguous circumstances such as body movements,

Fig. 1 Human activity model

multiple views and changes in environment. A numerous types of postures and activities have to be modeled to formulate an effective the recognition system. Human behaviour detection based on postures and activities need cautious attention as the human behaviour to a great extent depends on spatiotemporal factors in the environment as well as physiological characteristics of the human body such as velocity, acceleration, blood pressure and heart rate. This paper mainly focused on the postures and activities that are crucial for implementing an automated health caring system, such as detection of human falls under different situations. The model should cover all types of postures under various environments. Basic activity model is shown in Fig. 1.

Identify postures of unintentional falls

Human falls may occur while doing regular activities including standing, walking through the normal surface and a slippery surface, bending, trying to sit, trying to wake up from the bed, and so on. Some falls may lead to serious injury such as broken bones, and head injury which requires urgent medical attention, while some of them may not. Falls which need serious medication have to be identified based on the way the person is responding. An intelligent health caring system is ought to effectively identify the actual falls in real-time and proper information should be passed to the caretaker.

Identify postures of fall like activities

Some of the fall postures have a resemblance to normal postures such as bending, sitting, and lying on the floor. Also, postures of persons having neurodegenerative diseases such as Parkinson's disease may have a resemblance in normal activity and fall like activity. An effective person monitoring system has to differentiate postures of actual falls as well as fall-like activity similar to an actual fall.

Identify postures of normal activities

Daily life activities have to be modeled to segregate different movement patterns of the human body which is used to find out people with movement disabilities. By adjusting the body by spreading hands or catching on an object,

sometimes a genuine fall event becomes nonfall, particularly among individuals who are afraid of falls. Therefore careful attention is required to design an activity recognition model for implementing a real-time person monitoring system.

Identify other physiological factors

Brain-like activities are also highly influenced by physiological features. The applications that are controlled by these features could effectively monitor the person's physiological state. Many existing systems utilize physiological factors such as pulse rate and blood pressure as the measures to detect a fall. A sudden increase in these factors may represent a fall. But only using these factors, the detection system will not be accurate. Acceleration of person movement and its sudden changes can be measured along with the above-mentioned factors using different types of devices. Fall detection in bathroom areas where privacy is the primary concern, can utilize these physiological features effectively.

Design challenges and requirements

The fundamental challenge in brain-like computing in health-care field is the way the brain-like functions are copied to a computing model in order to build a system with decision making capability. Lack of labeled information makes it difficult to process and implement many real-time applications. In contrast to conventional learning methodologies, deep learning methods have the capability to generate features automatically which reduces the cost of feature engineering. Researches in data collection become a live topic in various communities including data processing and computer vision groups because of the significance of dealing with huge quantities of data. Along with normal human activities, an intelligent caretaking system should have the capability to detect anomalous behaviour such as a fall. A number of requirements to be satisfied such as accuracy, efficiency, cost-effectiveness, and real time monitoring. There are many difficulties to be faced to get an accurate image of a falling person as the presence of obstacles is more while capturing the image using a surveillance camera. The extraction of human body features from these images should be done carefully. Another problem is the difficulty in identifying various types of postures accurately as the recognition process needs to happen in real-time. Also, it is important to distinguish falling postures as well as normal human activity postures effectively. Otherwise, some normal postures may be wrongly identified as a person falling. So these types of false positives should not happen.

Camera-based methods raise privacy issues while considering fall detection in bathroom areas. Careful design is essential to overcome these types of problems. Person

Fig. 2 Posture detection methods

falling may happen in different situations such as while walking, standing, sitting, waking up from bed, or while doing other activities. Identification of falls during all these situations needs to be done. But this is not an easy task because the system needs to capture the postures under all these different circumstances. Along with this, complex scenes also need to be handled such as scenes capturing from different viewpoints as well as overlapping of images while tracking moving objects. Background conditions and complexity in lighting have also need to be considered. Also, a person falling can occur in different directions such as forward, backward, or in other directions. Postures for all these scenarios need to be evaluated. Under all these different circumstances, constructing an efficient fall detection system is essential, so that an automated person monitoring system will become successful.

In the cases of multiple persons residing in the same room, the activity recognition for all these persons has to be done in parallel. Ambiguous actions cause false interpretations since the same posture can have a different meaning in different situations. The system needs to distinguish fall activities as well as day to day activities. Fall-like activities also need to be considered since the chance for false prediction is more in these cases. While considering posture recognition in elderly people having some neurological diseases such as Parkinson's disease, the recognition process becomes more complex. PD patients are suffering from have muscular problems such as stiff muscles. Because of this reason, body movements and postures will be different and the chance of falling will be more.

Posture recognition system

As it is very hard to recognize the characteristics of a person falling, a careful study is required to design an efficient person monitoring system. A number of gait analysis and fall detection methods are available. Figure 2 shows the classification of posture recognition systems namely sensor-based methods and vision-based methods.

Sensor-based methods

The competence of a person monitoring system exceptionally relies on how efficiently the sensor detects the characteristics of the subject of interest. According to Glenn et al., sensors are categorized as passive or active. In passive sensors, the feature detection process directly depends on the presence of natural energy. Whereas active sensors have their own energy sources for irradiation. Fall identification frameworks have been structured utilizing either exterior sensors or wearable sensors.

Exterior sensors

Exterior sensors are installed nearer to the subject of interest. In [4], a person falling system implemented for the controlled environment has explained. The system uses passive infrared sensors that continuously measure body temperature. Radiated heat collected by the silicon lens converts the temperature into electromotive force which is used to calculate the temperature with an analog circuit. Performance of fall detection has measured through different recurrent neural networks and found that Bi-directional recurrent neural network provides excellent results compared with other methods. Guodong Feng et al. [39] proposed a method that uses pressure sensors embedded in the floor. This method has focused specifically on fall detection in bathroom areas. Fiber cables are placed on the floor. Pressure sensors associated with the fiber cables measures floor pressure which is used as an input to a BackProjection algorithm which reconstructs the pressure images and using that postures will be identified. A similar type of study has been done in [5] that uses strain gauge load cells located in smart tiles to measure the forces applied to the floor which tracks the position as well as various activities of a person. But this method is costly and also implementation is difficult inside the home. Shiba et al. [42] studied fall detection based on a Microwave Doppler sensor. Variations in velocities in the trajectory distinguish fall movements and non-fall movements based on Doppler Effect feature. Researches for real-time fall detection has successfully implemented using Wi-Fi devices [6]. A reliable segmentation has been done between fall and other similar type activities. The non-stationary behaviour of radar signals is utilized for distinguishing fall as well as non-fall activities [7]. Unobtrusive fall detection has been done using widened and transliterated variants of wavelet functions as the base. Wavelet transform coefficients are applied to a certain scale and form the feature vector at various scales over time.

Another study shows the implementation of activity recognition and fall detection system using wireless networks [8]. The system utilizes Channel state information (CSI) as the criterion for activity detection. Performance under various

scenarios in an indoor environment has been evaluated by utilizing 802.11n NIC. Activities of a single person located in a specific area have been implemented using this method. A similar type of technique has been suggested by Wang et al. [43] where the CSI speed-activity model is used to appraise the speed of movement of various parts of the human body and assorted human activities. A fall detection system based on tri-axis and PIR sensors are used in [9], where person detection has been done by calculating the temperature difference happens between two different infrared sources such as person and wall. When infrared sources with two different sensors cross each other, the varying voltage levels are stored. The presence of an accelerometer detects the speed levels of the moving object.

Wearable sensors

Wearable sensors are an option in contrast to exterior sensors. The sensors are implanted in wearable devices such as watches, shoes, belts, and buttons. It can also be placed in walking aid devices such as a walking stick, walker, and cane. Wearable sensors may be utilized for measuring motor associated symptoms including gait and stability analysis. These patient-friendly sensors can be easily positioned at different parts of the body. Using this, a caretaking system will be developed which continuously monitors affected persons' movements and provides clinically relevant data for further studies. As the technology has enhanced, wearable sensors had been advanced to quantitatively analyze movement patterns as well as day to day lives of the elderly. The sensor can be placed in the body so that the time and frequency domain features have to be extracted which will distinguish the gaits and postures of different types of people. Daily activities can be analyzed by extracting clinical data from Ground Reaction Sensors. Along with daily activity identification, an intelligent caretaking system should have the capability to identify different types of falls. Dharmitha et al. studied different sorts of falls such as forward falls, backward falls, sideward falls that happened while standing, and while trying to sit. Feature extraction has been done using a wearable sensor named MetaMotionR sensor. An edge computing framework based on a long short-term memory model categorizes the falls with an accuracy of 95.8 percentage.

Apply wavelet transforms to the data collected from a triaxial accelerometer and a pressure sensor [10] to detect sitting and standing postures as well as to segregate risky and non-risky falls. Studies show that the implementation of wavelet transforms in non-stationary signals of body movements provides outstanding time-frequency localization. Accelerometer-based fall detection has limitations to identify day-to-day activities. Along with accelerometer, Mao et al. [40] used a tri-axis gyroscope and a magnetometer to collect data. The orientation of the body has been calcu-

lated from Euler angle. The threshold-based fall detection system will be accurate only if the comparison is done with an exact threshold value of acceleration. The exact positioning of sensors in the body is also an important factor. Physiological measures such as pulse rate, blood pressure, and velocity are also being considered to assess the acuteness of fall. For different types of sensors, capabilities and sensing modalities are different.

Despite the fact that various physiological estimates should be thought of, wearable sensors offer a prevalent response for most by far of them [11]. Wearable sensors are reliable sources to capture body movements and other indispensable features of the human body. But these sensors require extra contemplation such as an easement, power requirement as well as the size of the gadget worn. In wearable gadgets, power prerequisites can immensely lessen a sensor's viability. Cates et al. [36] introduced an insole based system based on the signals from Force Sensitive Resistors (FSR) and Inertial Measurement Units to detect occurrences of falls through high and low acceleration activities.

Khanh et al. [37] explained a sensor-based gait analysis for Parkinson's patients. The wearable sensor consists of an accelerometer with three axes. Two sensors are placed in the upper part of the human body. Data from these sensors calculate the amount of flexure that happened. The method is suitable for static postures and its performance is degraded in a dynamic range such as walking. Also biasing will be difficult in varying temperatures and voltage that may lead to inaccurate results. The integration of multiple sensors placed on various locations of the body makes the prediction more accurate. MPU6050 is a motion-tracking device consisting of a gyroscope and accelerometer, used to measure the acceleration and other features of the body [12]. Along with this, a graphene/rubber sensor can also be used to detect leg movements. Compared to conventional motion sensors, the MPU6050 sensor has low power consumption as well as an anti-inference capability which makes explicit motion detection in real-time. Tilt angles of the body and acceleration are directly proportional to each other. Abrupt changes in the speed of body movement and angle between the floor plane and the human body can be used as significant gauges for checking fall activity. Signal magnitude vectors indicate the intense movements of the body. Flexible sensors based on polyvinylidene fluoride attached in patient cloths measure the signals when bending or other types of movements happening in the body [13].

Hybrid sensors

Hybrid sensors are the combination of exterior as well as wearable sensors. The problems with camera-based methods are the chances of occlusion, ambiguities due to varying lighting conditions, and rapid movements of the body, leading to

decision-making more complex. Hybrid sensors incorporate inertial sensors with wearable sensors so that abrupt changes happening in the human body will be precisely estimated. This will make the decision making more accurate in varying environments. The usage of hybrid sensors solves the privacy issues related to wearable sensor-based methods. In this method, privacy issues will be solved with the help of exterior sensors. Actions under normal environments will be detected using efficient wearable sensors. The combination of both types of sensors makes the detection more accurate.

Vision based methods

A vision-based system makes use of visual aids such as video cameras for activity recognition. Intelligent health caring is a complex problem since one of the main tasks here is identifying various gaits and postures of human beings. This task is more complex in the case of aged people having neurodegenerative diseases such as Parkinson's disease. The postures of these patients are different from the postures of a normal human being. Normal activities performed by these patients may be mispredicted as some other activity. Therefore the normal, as well as fall activities, should be precisely estimated. In a real environment, many situations need to be handled, including changes in the environment, lighting conditions, etc. The main task in intelligent health caring is to address these types of complex circumstances, different postures, and gaits in real-time. A carefully designed system is essential to extend the applicability of the posture and activity recognition process.

With single camera method

Fall recognitions utilizing a single RGB camera is more economical than an RGB-depth camera. But RGB-D camera has the advantage of capturing depth information of the subject of interest. Also, depth camera will provide multiple views of the object in real-time. Body joint information is the most important data in a 3D body modeling system. Depth camera has the advantage of providing joint information about the human body. A 3D point cloud method of posture and fall identification system has proposed by Peng et al., [41] where the depth information was captured using a Kinect camera. A combination of changes in acceleration and body height are the features considered for fall detection. Based on a fall hypothesis, whether it is true or false, decides the orientation and position of the person as for the floor position. Kinect camera with a pulse coupled neural network suggested efficient posture identification in a littered environment [14]. Usage of a depth camera for capturing images resolves the multiple viewpoint problems associated with the traditional 2D camera. A pulse coupled neural network with an image sensor acts as a harmonious interface with the processor.

In threshold-based fall detection, feature extraction and classifications processes can be avoided by utilizing the parameters such as human centroid and edge between the human body and floor surface [15]. Extraction of a person's shape features while moving can be calculated using moment functions and are portrayed by an ellipse shape. The decision to choose a basic model is the pivotal step in every recognition system as the quantity and the type of information to be represented are varying model by model. For example, dominant postures sitting, standing, falling, and lying down are described with different joint parameters of the body. Researchers proposed several 3D body modeling techniques for posture recognition. Many of them are fully or partially successful in their research. In a 3D model, the main parts to be represented for posture recognition and fall detection are the human body's head and leg positions. 3D surface plots the structure of the head as 3D points [16]. A person's position with respect to an object such as a chair is examined using the YOLO object detection algorithm. Camera-based methods raise privacy issues while considering fall detection in bathroom areas.

With multi camera method

The main problem associated with a single camera method is the chance of occlusion. In a real-time person monitoring system, sometimes it won't be easy to view an occluded object with one camera. This problem can be solved using a multi-camera system that captures numerous images of the same object from various perspectives. That is, another camera can view an occluded object. This approach's general objective is to monitor all the objects irrespective of the environment in which the process has to be done. It would be essential to perform data integration with multiple sensors just as between the items seen by a single sensor. Espinosa et al. proposed a technique for fall identification by using multiple cameras in conjunction with CNN. The system excels in performance compared with traditional machine learning methods and has the ability to compete with the state of art methods. But privacy issues and environmental changes are not addressed in this method.

Stages in human action recognition

The steps to identify falls from various daily life activities are shown in Fig. 3.

The activity recognition process consists mainly of two stages: the activity learning phase and the activity recognition phase. The activity learning phase focused on building a well-structured activity learning model. The formation of the model has undergone several stages. The constructed activ-

Fig. 3 Activity recognition stages

ity model has applied to recognize the activities of a person from real-time videos.

Capturing videos

Recognition of human postures is a challenging task when considering the processing of video sequences. Several classifiers are required for training and testing process. The recognition rate is the most important matrix that needs to be satisfied. The recognition rate highly dependent on the number of postures trained and tested. Video capturing based on single camera and multi-camera methods are discussed in the previous sections. The latest method of capturing videos is based on a camera with RGB-depth sensor. This consists of a camera with a depth-sensing device. RGB-D data set can be constructed by using Kinect camera [17]. All types of postures, including static and dynamic situations, are considered for posture recognition. Images are captured with different distances and orientations. So scalability and orientation features can also be evaluated. It provides RGB images with depth information. That means it gives some information about the distance between the sensor and the object that need to be captured on a pixel by pixel basis. The shift in the dotted pattern identifies the depth of a particular region. It captures the images with some depth information such as distance between the sensor and the object on a pixel by pixel basis.

Preprocessing and feature extraction

With the arrival of high-speed machines with enormous storage capacity, many kinds of research are happening to process and analyze a massive amount of data. It is challenging to deal with huge data quantity in real-time applications, mainly in decision-making problems. For making better decisions, several variables need to be considered. Preprocessing is mainly focused on improving the quality of captured data before start actual processing. This includes removing some unwanted data, enhancing a suppressed feature, and so on for further processing. Various filters are used either to smoothen or to enhance an image. Smoothening refers to the conversion of an image with high frequency to an image with low frequency. Removal of noise present in the image can be done using mean or median filtering techniques. Image smoothening has been performed with Gaussian function, based on convolution operation with the 2-D distribution. A blurred image will be formed as a result of smoothening. In the case of object tracking, edge detection is an important task where the presence of an edge will be detected based on the expeditious changes in the image's intensity. The rapid change in the intensity has to be identified using Laplacian filters. The selection of proper data preprocessing technique results in creating a more appropriate set of information suitable for feature extraction and activity prediction.

Learning from a huge amount of data is not an easy process. This feature extraction process is known as dimensionality reduction. This is the process of retrieving only relevant attributes that enclose the entire data's meaning without loss of generality. One of the most commonly used dimensionality reduction technique is Principal Component Analysis. In PCA, find a combination of features from the original dimension, which summarizes the given data in the best suitable way. Thus, the pairwise distance between all the features is calculated by maximizing the variance and minimizing the error. Given data is represented as orthogonal axes, and each of them is ranked according to the importance. Dimensionality reduction based on Linear Discriminant Analysis maximizes the mean distance between each class as well as the class itself. Classification accuracy will be improved by increasing the distance between the mean of each class. LDA is based on Gaussian distribution for inputting data, which increases the accuracy of image classification. While in the case of Locally Linear Embedding, the process of feature extraction depends on Manifold Learning. A Manifold represents an object in 3D dimensional space, embedded in higher-dimensional space. Nonlinear features are extracted by applying LLE to the approximation components so that time-frequency features are obtained from the components [18].

Segmentation

Image segmentation

Image segmentation is used to split the entire image into parts dependent on image properties, such as texture, color, orientation, and intensity [19]. Segmentation also makes the processing of the videos simple, and it helps to easily extract the nutshell of the video. Segmentation based on threshold uses pixel intensity values as the key feature for segmentation. Spatial requirements are not considered during segmentation. The proposed system uses threshold-based segmentation to perform foreground and background segregation. In this method, two categories of images will be formed depending on the pixels' intensity values. Pixel values between a certain range will be considered those forms of foreground objects, while the other range of values forms the image's background. This method is used when the brightness of the object and the background are different. Spatial requirements are not considered during segmentation. If the pixel value passes the threshold test, then the corresponding pixel is considered a part of the foreground object; otherwise, it is considered part of the background image. The resultant image $p(x,y)$ will be calculated from the original image $q(x,y)$ using the following method.

$$p(x, y) = \begin{cases} 1 & \text{if } q(x, y) > \text{Threshold} \\ 0 & \text{if } q(x, y) \leq \text{Threshold} \end{cases} \quad (1)$$

In edge-based segmentation, detected boundaries are connected to form the object boundary. Using this, the required regions will be segmented. The most commonly used edge-based segmentation is Gradient-based segmentation. This method is not suitable if the number of edges in the object is more. In segmentation based on region growing, a region with similar characteristics is formed by combining adjacent pixels. It is expensive in terms of time and space. Segmentation based on clustering uses a membership function with values either 0 or 1. These values represent the specific cluster in which a pixel belongs. We can use either k-means or Fuzzy C-means clustering. This is useful in real-time applications, even if the method to find a membership function is difficult.

Video segmentation

Video segmentation means splitting the video into meaningful shots [20]. Video signals consisting of temporal information such as camera movements, object movements, and so on. Various video segmentation techniques are given in Fig. 4:

Fig. 4 Video segmentation

Background subtraction The main task in object recognition is the extraction of foreground objects by removing the background details. This process is crucial in the case of real-time video surveillance systems [21]. One of the most important types of video segmentation method is background subtraction or frame differencing [22]. Initially, measure the background for time t . Then find the difference between the current frame and measured background. The foreground mask will be estimated by applying a threshold value to the absolute difference. Image at a certain time t , $Im(x, y, t)$ and background $Bg(x, y, t)$ are measured. Calculate the absolute value as:

$$Ab(x, y, t) = Ab(Im(x, y, t) - Bg(x, y, t)) \quad (2)$$

If $Ab(x, y, t)$ exceeds the threshold, then output Ab as a result.

Kernel based FCM Fuzzy C-means algorithm is modified by adding some kernel information to overcome the problems in FCM [23]. It can also handle clusters that do not have a considerable difference. In this method, lower-dimensional space features can be mapped to higher-dimensional space, also called kernel space. This method can handle both linear and non-linear features of video sequences.

FCM based on histogram In this segmentation, changes in shots have to be identified [24]. Apply the clustering algorithm with respect to the changes in the shots. Differences in the color histogram are considered as the features extracted for clustering. A fuzzy C-means algorithm is used to form various clusters based on the variations calculated from the feature extraction step. Frame selection has been made in a heuristic approach. Video sequences will be segmented into different shots. Find the centroid frames from each shot, and these frames are chosen as the keyframes. FCM segmentation based on cellular automata uses image features weight and spatial feature weight [25]. Cellular automata utilize a dynamic model where an object's discrete features are used for segmentation. The discrete features of an object include state, time, and space. A complex dynamic system has been effectively simulated by applying its own parallel computing capability in this method. The cellular automata model

Fig. 5 CNN window

extracts the intended area by extending the binary image to a grayscale format. Along with this, the model is also expanded to multiple value state. The classification accuracy has been improved by applying continuous modification to the cell located in the center.

Object detection and activity model formation

Object detection involves mainly two tasks: classification of image and localization of objects under consideration. Image classification involves identifying each object's class in an image, and based on that, each image is assigned a class label. Image classification outputs integers mapped to the assigned class labels. Image classification algorithms are designed to output the presence of different objects in the input image. The most commonly used image classification model is Convolutional Neural Network (CNN). Using fewer parameters in CNN reduces the complexity of the training process. The amount of data required for learning a model has also been minimized. CNN works by looking at small portions of an image, which gradually leads to learn and process the whole image.

Figure 5 shows a sample CNN window where an image is scanned as 4×4 windows. The window movement is based on stride length. The movement will be completed once the entire image is scanned. The weighted sum of the pixel values of each window is calculated. Image classification is not enough for getting a complete understanding of the image [26]. The accurate analysis of the objects' concepts and position is also essential in object detection because the orientation, lighting conditions, and viewpoints may vary in different environments. In object localization, identify each object's location in an image and draw a boundary box around each of them. Objects in the image will be located, and the class of the image will be identified. Bounding box regression utilizes multi-feature combination and multi-scale adaption properties for object recognition.

R-CNN is a simple and straightforward application of CNN for object localization and recognition. In R-CNN, training is a multistage pipeline. Therefore it is expensive while considering time and space. For object localization and segmentation, highly capable CNN can be applied to bottom-up region proposals. This is a supervised pre-training paradigm which reduces the problems in vision data sets, whereas in Fast R-CNN, a convolutional feature map is generated by inputting the image directly to the CNN. It requires a set of candidate regions along with each input image. Region of proposal will be identified, and reshaping has been done for making the result to feed into a fully connected layer. This is used to classify the objects in the proposed region in an image [27]. State of the art object detection can be efficiently implemented with Faster R-CNN. YOLO v1 uses a unified model for object recognition in real-time applications. YOLO family of models is much faster than R-CNN that provides object detection in real-time [28]. YOLO model identifies the objects present in an image as well as the position of objects easily [29].

Advantages

1. Easy to construct
2. Training can be done directly on a complete image

YOLO version 2 is an advanced model for object recognition where bounding box coordinates are formed directly from image pixels, which improves the classification probability. YOLO v2 is based on a technique known as batch normalization, where the value distribution is normalized before going through the next layer. This improves the accuracy of object detection in a short period of time. A finely tuned classification network will be formed using 448×448 images for 10 epochs on ImageNet. Instead of fully connected layers, YOLO v2 uses anchor boxes for boundary box prediction. YOLO v3 is a highly accurate and faster model for object detection [30]. Dimension clusters are used as anchor boxes for the prediction of bounding boxes. Four coordinates are considered for each bounding box. Bounding box properties are calculated using logistic regression. Logistic classifiers are used for predicting the class of bounding boxes. Salient object detection methods use pixel-level segmentation and contrast enhancement for object detection. Performance of various object detection models is shown in Fig. 6.

Activity recognition and model evaluation

Human action recognition can be done from video sequences based on 3D coordinate systems in conjunction with Hidden Markov Models, especially Cyclic HMMs [31]. 3D human modeling techniques identify various joints of the human body through continuous cycles. Apply dimensionality reduction to these coordinates and then k-means clustering

Fig. 6 Performance comparison of various object detection models

to generate feature vectors in the form of clusters. These vectors are used to train Cyclic HMMs so that different types of actions can be identified. Human postures are captured and analyzed using the Kinect sensor [32]. Silhouette information is separated from the background using the background subtraction method. Various image detection methods are combined to form a hybrid method that is used for posture recognition. A Kinect sensor has the following components:

1. A sensor to detect the depth information
2. RGB camera
3. Motorized tilt to adjust sensor's angle

Depth image means the distance between the sensor and the object needs to be identified. The sensor cannot give any lighting and color information available in a depth image. In this method, the center of gravity of the silhouette is calculated, and based on that; the silhouette is split into upper and lower parts. The width of the body is also calculated by projecting the histogram horizontally. Using this, the distance between the silhouette edges and the center of gravity can also be calculated. Different curves in the silhouette are obtained from this method. The linear Vector Quantization method is used for the classification of different postures. The advantage of this method is that the recognition rate is high and can be done within a short period of time. Also, this requires only a limited data set. As the distance from the sensor is increased, the size of the object will appear as smaller. Therefore the accuracy of the posture may be affected. Some form of normalization needs to be performed to avoid this problem.

Support Vector Machine is utilized to classify various postures such as standing, crouching, walking, and lying. Results are analyzed with various types of classifiers. Among them, quadratic SVM shows better performance compared with other classifiers. The quality of depth images decreases as the distance from the camera increases. Therefore the quality of the silhouette segmentation will be reduced. The performance of this method is suitable in a closed environment.

Also, performance is varying in varying lighting conditions. Postures can be identified from video sequences by combining 2D methods with three-dimensional models of the human body [33]. Human posture identification is implemented in two ways. One method is based on tracking the markers of the body, and the other one is using vision-based algorithms. The second method is more suitable for people who are not cooperating with these types of recognition activities. From the video, images are captured. A binary image is obtained by measuring the difference between the current image and the reference image. Various features, including 3D position, height, and width are calculated. The 3D human silhouette is constructed initially by using a 3D scene generator. Various parameters needed for a specific posture will be given as the input to this generator. Mesa graphics library is used to construct a 3D human model. A set of 23 parameters are utilized for human posture representation.

Four types of general postures are studied, such as sitting, standing, bending, and lying. A human 3D model can be formed by applying all these parameters at different positions. In a scene, the visibility of objects will be increased using a Z-buffer. This is used to provide depth information between the camera and the object need to be detected. The pixel value corresponding to the z -coordinate in a 3D scene will be represented in an XY plane. Based on depth information, each time, this value will be updated when the image is closer to the camera. The old values of a pixel will be hidden behind the new values. The inclination of the model is done by considering all types of rotations in the scene. A person facing the camera is considered as zero degrees orientation.

The type of posture decides the position and axis of rotation. Once the silhouette is obtained, it can be described with various factors such as Hu moments, skeletonization, etc. Euclidean distance is calculated, and the least distance posture will be considered a more accurate posture. The similarity is calculated by finding the distance between adjacent vectors. Equation 3 shows the angle between two vectors **A** and **B**. Positive angles are known as an acute angle; otherwise, it is known as an obtuse angle.

$$\cos \theta = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} \quad (3)$$

The resultant similarity value is applied in the distance formula with two vectors $V1$, and $V2$ corresponding to x and y coordinates are given below.

$$\text{Distance}(V1_{xy}, V2_{xy}) = \sqrt{2 \cdot \cos \theta (V1_{xy}, V2_{xy})} \quad (4)$$

In skeletonization, silhouette contour is extracted by using a method known as thinning. The silhouette is enlarged, and various anomalies were removed, and the border is extracted. Statistical moments are used to identify the centroid of the

silhouette. Euclidean distance between detected posture the model posture will be calculated. The one which minimizes the distance gives an accurate result. Different postures may form ambiguity in image projection at certain viewpoints. To avoid that, synthetic data can be used to detect ambiguous postures. Human action recognition can be done from video sequences based on 3D coordinate systems in conjunction with Hidden Markov Models, especially Cyclic HMMs [31]. 3D human modeling techniques identify various joints of the human body through continuous cycles. Apply dimensionality reduction to these coordinates and then k -means clustering to generate feature vectors in the form of clusters. These vectors are used to train Cyclic HMMs so that different types of actions can be identified. Separately trained models are combined to recognize continuous actions.

Experimental setup

Previously, human object detection has in the form of drawing bounding boxes around the objects. But the latest technologies have included many more features, including detection of bounding masks associated with the object and projecting various key points of the human body in real time.

Pose estimation using detectron2 model

Detectron2 is a platform for implementing these features, especially keypoint detection and labeling various parts of the body in an efficient manner [34]. It detects and labels each of the objects in the image, including background objects. The bounding box has an extended feature, such as bounding the object with a colored mask, which gives the complete structure of the object under consideration. That means detectron2 provides a complete understanding of all the objects included in a particular scene. Compared to the earlier version of detectron, detectron2 is an open-source framework with a faster and accurate model for object detection and pose estimation. One of the key features of detectron2 is its modular design. Training efficiency is improved in detectron2, and it is completely built under the PyTorch platform.

Background separation and object detection

The workflow of Detectron2 based background segregation process is shown in Fig. 7. A binary image will be formed by applying some threshold value to the existing grayscale image. The Black and white representation of an image is created by considering the threshold value. The threshold values can be chosen either manually or applying algorithms known as automatic thresholding. Once the threshold is calculated, each pixel value in the image is compared with the threshold value. If the threshold value is greater than the pixel value, the

Fig. 7 Workflow of background separation process

current pixel is replaced with a black pixel. Otherwise, the current pixel is replaced with a white pixel, and this process will be applied for the entire image to form the equivalent grayscale image. The feature map calculation has been done using the following equation:

$$F(m, n) = (i * k)(m, n) = \sum_p \sum_q i(p, q)k[m-p, n-q] \quad (5)$$

Equation 5 depicts the feature map calculation to perform background subtraction, where i is an array of pixels, and k represent the chosen kernel. The number of channels ranges from p to q and flipping of the filter is denoted as $k(m-p, n-q)$. If the stride length is large, then the spatial dimensions corresponding to the feature map can be reduced [35].

To provide a transition to a 3D space, apply multiple filters to the same image. The convolution process will be performed each of these filters independently and combine them to form the result. The dimensions should ensure the following equation.

$$[n, n, c] * [f, f, c] = \left[\left\lfloor \frac{n+2p-f}{s} + 1 \right\rfloor, \left\lfloor \frac{n+2p-f}{s} + 1 \right\rfloor, l \right] \quad (6)$$

Here n denotes the size of the image and f denotes the size of the filter used. Number of channels are denoted by c . p and s provide padding and stride information. l denotes number of filters used in the transition process. Various layers during the translation process are shown in Fig. 9

The Black and white representation of the mask is multiplied with the input image matrix. The result is used to form the masked representation of foreground objects. The obscured input image is multiplied with mask value to form a masked representation of background objects. The extracted masked foreground and background are added to form the

resultant output image, which is focused only on the objects of interest. Detectron2 repository has a large number of classes, including pre trained models and many other necessary classes. A trained detectron2 model has been used for object detection. In this work, the person detection model utilizes images from the COCO image data set. Real-time human poses are captured with a normal digital camera and given as input for the object detection process. The images with varying poses are shown in Fig. 8. Images of poses such as walking, sitting, bending, lying on the floor, trying to rise from floor, bed, and chair are captured. The Black and white representation of the mask from the input image is shown in Fig. 10. Extracted foreground and background separated images are multiplied with the mask value to form exact output images representing various poses of the human body from Fig. 8. As this work aims to identify various human poses, even though several objects are present in the image, the segmentation process is focused only on person objects.

$$\begin{bmatrix} \text{False} & \text{False} & \text{False} & \dots & \text{False} & \text{False} & \text{False} \\ \text{False} & \text{False} & \text{False} & \dots & \text{False} & \text{False} & \text{False} \\ \text{False} & \text{False} & \text{False} & \dots & \text{False} & \text{False} & \text{False} \\ \dots & & & & & & \\ \text{False} & \text{False} & \text{False} & \dots & \text{False} & \text{False} & \text{False} \\ \text{False} & \text{False} & \text{False} & \dots & \text{False} & \text{False} & \text{False} \\ \text{False} & \text{False} & \text{False} & \dots & \text{False} & \text{False} & \text{False} \end{bmatrix}$$

[128.61908 17.961678 181.85457 120.99814]

The matrix shows the mask representation of an image in the form of true and false. The matrix value tells where the pixels in the large image are true for the mask. The coordinate values represent the upper left and lower right corners of the boundary box associated with one object. The bounding box coordinates are used to calculate the object's height and width, which gives the exact position and size of the object. Figure 10 shows the black and white representation of a mask associated with the identified person object and the pixel values in the input image. The width and height of the person object is exactly calculated using the bounding box coordinates mentioned above. The result shows that detectron2 platform gives better accuracy compared with existing object detection algorithms. The same approach is used to detect human key points from real time videos. In this case, the key point localization of person movements has to be done. This is an added feature in detectron2 to locate the person object and postures exactly. Figure 11 shows the output obtained after processing the raw input image. Detected objects with a bounding box and associated accuracy are shown in the figure. The objects are completely covered with a mask with different colors, which gives the object's exact structure. Predictions on the image have been made with the help of a visualizer function. The helper function makes the image

Fig. 8 Real-time human poses: (a) Bending, (b) walking, (c) bending, (d) fall position, (e, l) Sitting, (f) trying to rise up after fall, (g) Fall position (h, j, k) Get up from bed (i) Sitting

Fig. 9 Layers in transition process

comes inside the mask. The calculated height and width are used to make the image inside the mask.

Keypoint estimation

The same real-time images are shown in Fig. 8 are used as sample images for keypoint detection. From detectron2 various files are imported, such as logger, visualizer, metadata, and zoo models. Zoo models provided weights for processing the video. Supporting configuration files are included adding configuration function. Detectron2 provides all this information as detectron2 utility files. Model and its waits are passed as arguments in the file. The threshold value to this model is made as 0.7. The default predictor predicts by utilizing the model and its threshold value. Forward propagation is implemented by passing the original image vector to the default predictor and then to the specified model. As it is difficult to implement forward propagation using a graphics processing unit (GPU), instances are converted into CPU so that the problems to produce output images were solved.

This work includes 17 important keypoints of the human body shown in Fig. 12. These points are used for pose estimation and activity detection. The following equation is used to

reduce the effect of lower confidence values in distance measurements between the keypoints:

$$\text{Distance}(V1, V2) = \frac{1}{\sum_{i=1}^{17} V1_{c_i}} \times \sum_{i=1}^{17} V1_{c_i} \|V1_{xy_i} - V2_{xy_i}\| \quad (7)$$

Out of 17 keypoints considered, for each keypoint i , $V1$ and $V2$ are the vectors corresponding to x and y coordinates. $V1_{c_i}$ is the corresponding confidence value. The metric used to measure the accuracy of keypoint detection in COCO data set with scale factor s , a positive integer k_i , and visibility value v_i is given below:

$$\text{Similarity} = \frac{\sum_i e^{-\frac{d_i^2}{2s^2k_i^2}} \delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \quad (8)$$

Here, d_i denotes the distance between actual value and predicted value.

All the frameworks with dependency issues were rectified so that results become more accurate. Bounding box representation decides the exact location of each object in the image. Detected key points and the associated accuracy from various human poses are shown in Fig. 13. The key points are formed by identifying various joints and other important points of interest in the human body. The key points include all joint positions, eye, ear, and so on. By joining all these points, the body's structure will be formed, and this joined keypoints identifies the poses at a particular time. The method's main advantage is that it does not require high-end devices for capturing images and videos. Key point detection of poses such as bending, walking, standing, falling, sitting, lying on the floor, and rising from the bed is identified successfully.

Performance evaluation

The evaluation metrics of keypoint detection are precision and recall, which gives a homomorphism measure between

Fig. 10 Black and white representation of mask

Fig. 11 Bounding box and mask representation

actual truth value and predicted value. The two measures are calculated as follows:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (9)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (10)$$

Fig. 12 17 keypoints of the human body

Precision shows all truth values, including true positives and false positives, how many are actually positive. Where recall says information about positive predictions that are missed due to some factors. Table 1 shows the evaluation metrics and resultant values for keypoint detection process. The accuracy of the keypoint detection is measured with two metrics, namely average precision and average recall. The values of these two measures range from 0 to 1. Let $p(x)$ denotes the precision of the keypoint detection calculated using the above equation. Then the average precision is calculated as follows:

$$\text{Average Precision} = \int_0^1 p(x) dx \quad (11)$$

Performance evaluation of the detectron2 model for key-point detection has been done by considering different types of metrics. Intersection over union (IoU) is a metric used to measure the boundary overlapping between ground-truth value and predicted value. This metric provides better results in the case of predicted bounding boxes. IoU is the ratio of the area of overlap and area of union. This metric is recom-

Table 1 Evaluation results

Metrics	Boundary measure	Maximum detection	Area	Value
Average precision	Intersection over union	100	All (AP)	0.683
		100	All (AP50)	0.972
		100	All (AP75)	0.768
		100	Small (APs)	0.269
		100	Medium (APm)	0.704
		100	Large (API)	0.502
Average recall	Intersection over union	1	All (AP)	0.483
		10	All (AP50)	0.724
		100	All (AP75)	0.775
		100	Small (APs)	0.348
		100	Medium (APm)	0.899
		100	Large (API)	0.521

Fig. 13 Human pose keypoint detection

mended since unrealistic outputs are obtained due to varying window size and feature extraction methods. Average precision and recall have been calculated over different areas such as small, medium, and large. The measure *maximum detection* gives maximum precision or recall given the number of

Fig. 14 Performance evaluation

detections in an image. AP50 and AP75 give the ground truth values over 50 and 75% of the regions, respectively. Performance metrics for bounding box evaluation in keypoint detection is shown in Fig. 14. From the graph, it is clear that AP50 has maximum detection capability compared with other performance metrics. Average precision over small, medium, and large size objects are denoted as APs, APm, and API, respectively.

Object detection and activity recognition using PoseNet model

PoseNet is a pre-trained opensource machine learning model for human pose estimation. This model easily tracks the human body in real-time with a webcam image and provides output as the key points associated with the body in the form of x and y coordinates. These coordinates provide the confidence value of each pose and form a skeleton of the body. Pose property utilizes the 17 key points present in

Fig. 15 PoseNet model for activity recognition

the human body. The model is trained using COCO data set, which consists of a large number of labelled images. As the PoseNet model works on event handlers, turn the pose event on. The model detects multiple poses at a time if there is more than one person present in the image. Figure 15 shows the PoseNet model implemented for classifying normal and abnormal activities. The captured image is passed as an input to the PoseNet model, which guesses the exact positions of all the key points and classify the image as normal and abnormal.

PoseNet works under both single pose and multi-pose detection algorithms. Tests have been made by comparing MobileNetV1 and ResNet50. Performance of single pose and multipose detection in MobileNetV1 and ResNet50 are shown in Table 2. Image resolution positively affects the performance of the system. Higher the image scale results in higher precision yet lower speed. The resultant confidence values are shown in Fig. 18. Images are processed with different stride values. The feature map size based on the input image size is calculated using this feature. If the output stride value is 16, the feature vector's size will be 16 times smaller than the input image size. As the stride value increases, the performance is increased while accuracy got reduced. The multiplier feature directly affects the performance of the system. As the multiplier value increases, the number of parameters used will also increase, which will produce better results, but the system's performance will be reduced in terms of speed. On the other hand, if the multiplier value is too low, the results will become worse. Therefore, in the proposed method, a moderate value of 0.75 is assigned to the multiplier. MobileNet weight quantization is affected by the feature quantBytes, an optional parameter. As the quantBytes value increases, the size of the model also increases. This increases the time for loading the model for training and testing. The proposed system assigns the quantBytes value as 2 in MobileNetV1 and 1 in ResNetV50. Frame rate represents the number of frames exhibited on display. It is expressed in Frames Per Second(FPS). The performance cost will be increased as the frame rate increases. The comprehensive estimation of confidences are done using PoseConfidence value. It may be utilized to conceal poses that are not adequate for pose estimation model. Figure 16 shows the results obtained under varying lighting conditions. The speed of

Fig. 16 Human pose detection under varying lighting conditions

Fig. 17 Measured framerate using MobileNetV1 and ResNet50

detection for both single pose and multi-pose algorithms by setting the stride values as 8 and 16 is shown in Fig. 17. Here the model is trained with full-body images. The distance between the webcam and the object is around 10 m. If the camera does not completely capture the body, then the accuracy may be reduced. Initially, tests are conducted for categorizing normal activities and abnormal activities. The network structure is determined by using a loss value hyperparameter. The number of hidden units is determined by using this value. Figure 19 shows the labelled output obtained after applying the model in two real-time poses. The tests conducted in a closed room with a single person present in the image. The skeleton property associated with PoseNet model is used to draw lines between various key points. The PoseNet output is given as an input to an ML5 neural network, which provides the appropriate labels for each pose. All the x and y coordinates are labelled accordingly. Using this, the classifier is trained to make an accurate guess on various activities.

Table 2 Performance of single pose and multipose detection in MobileNetV1 and ResNet50

Algorithm	Architecture	Input resolution	Output stride	Multiplier	Quantization (in bytes)	Frame rate	Pose confidence
SinglePose	MobileNetV1	200	8/16	0.75	2	11 FPS	0.88
		250	8/16	0.75	2	8 FPS	0.86
		300	8/16	0.75	2	6 FPS	0.85
		350	8/16	0.75	2	4 FPS	0.77
		400	8/16	0.75	2	3 FPS	0.75
		450	8/16	0.75	2	2 FPS	0.75
		500	8/16	0.75	2	2 FPS	0.73
		200	16/32	0.75	2	11 FPS	0.79
		250	16/32	0.75	2	8 FPS	0.78
		300	16/32	0.75	2	6 FPS	0.77
MultiPose	MobileNetV1	350	16/32	0.75	2	4 FPS	0.74
		400	16/32	0.75	2	3 FPS	0.73
		450	16/32	0.75	2	2 FPS	0.72
		500	16/32	0.75	2	2 FPS	0.7
		200	32	0.75	1	11 FPS	0.87
		250	32	0.75	1	8 FPS	0.84
		300	32	0.75	1	6 FPS	0.83
		350	32	0.75	1	4 FPS	0.8
		200	32	0.75	1	11 FPS	0.65
		250	32	0.75	1	8 FPS	0.64
SinglePose	ResNet50	300	32	0.75	1	6 FPS	0.64
		350	32	0.75	1	4 FPS	0.62
MultiPose	ResNet50	200	32	0.75	1	4 FPS	0.62
		250	32	0.75	1	4 FPS	0.62
		300	32	0.75	1	4 FPS	0.62
		350	32	0.75	1	4 FPS	0.62
		200	32	0.75	1	4 FPS	0.62
		250	32	0.75	1	4 FPS	0.62
		300	32	0.75	1	4 FPS	0.62
		350	32	0.75	1	4 FPS	0.62
		200	32	0.75	1	4 FPS	0.62
		250	32	0.75	1	4 FPS	0.62

Fig. 18 Performance of PoseNet models

Loss function minimizes the model error and the value obtained from this function is referred to as loss value. The proposed model utilizes the cost function as binary cross-entropy as two classes of activities have to be identified. Equation 12 depicts the process of calculating the loss value in a binary classification. The similar procedure applied in a multi-class category is also shown in Eq. 13. Here the optimized loss value is obtained by calculating the gradient of resultant neurons.

$$f(p_i) = \frac{1}{1 + e^{-(p_i)}} \quad (12)$$

Here p_i denotes the convolutional value used in the network

A similar procedure is applied to identify and label multi-class activities such as sitting, standing, falling, and lying on the floor. The resultant CNN vector is a multi-dimensional vector of positive classes C . The loss function applied in this classification is given below:

$$\text{Lossvalue} = -\log \left(\frac{e^{p_m}}{\sum_j^C e^{p_j}} \right) \quad (13)$$

Here the convolutional value of positive classes is denoted by p_m . Optimized loss value is obtained by calculating the gradient of resultant neurons. Following equations provide the derivatives of both positive and negative classes:

$$\frac{\partial}{\partial p_m} \left(-\log \left(\frac{e^{p_m}}{\sum_j^C e^{p_j}} \right) \right) = \left(\frac{e^{p_m}}{\sum_j^C e^{p_j}} - 1 \right) \quad (14)$$

$$\frac{\partial}{\partial p_n} \left(-\log \left(\frac{e^{p_m}}{\sum_j^C e^{p_j}} \right) \right) = \left(\frac{e^{p_n}}{\sum_j^C e^{p_j}} \right) \quad (15)$$

p_m and p_n denote the convolutional values for positive and negative classes. In this research, real-time pose estimation models were applied for object detection and abnormal activity recognition with vision-based complex key point analysis.

Fig. 19 Detected poses: (a) normal pose, (b) abnormal pose

Fig. 20 Falls in different directions

The proposed system implemented a state of the art computing model for the development of public healthcare industry. The normal and abnormal activities are identified, and the results are shown in Fig. 19. The comprehensive estimation of confidences has done using PoseConfidence value. It may be utilized to conceal poses that are not adequate for pose estimation model. PoseConfidence value can be used to filter out the responses that could not make any feedback in the pose estimation process. A higher value of PoseConfidence makes better decisions and will be more reliable. The PoseConfidence values obtained for single-pose and multi-pose estimations are shown in Table 2 with a maximum PoseConfidence value as 0.88 in a scale of 0–1. While the multiplier feature directly affects the performance of the system. As the value of the multiplier increases, the number of parameters used will also increase. This will output better results,

Fig. 21 Loss value for binary classification

but the performance of the system will be reduced in terms of speed. On the other hand, if the multiplier value is too low, the results will become worse. MobileNet weight quantization is affected by the feature quantBytes, an optional parameter. As the quantBytes value increases, the size of the model also increases. This will increase the time for loading the model for training and testing. The experimental results show that the models have high accuracy levels for detecting sudden changes in movements under varying environments (Fig. 20).

$$\text{Loss} = -\log(f(p_1)) \text{ if } y_1 = 1 \quad (16)$$

$$\text{Loss} = -\log(1 - f(p_1)) \text{ if } y_1 = 0 \quad (17)$$

Here $y_1 = 1$ means normal class and $y_1 = 0$ means abnormal class. When $p_i = p_1$ the corresponding gradient value is written as:

$$\frac{\partial}{\partial p_i} (\text{CE}(f(p_i) = y_1(f(p_1) - 1) + (1 - y_1)f(p_1)) \quad (18)$$

The loss values calculated using sigmoid function is given below.

$$\frac{\partial}{\partial p_i} (\text{CE}(f(p_i) = \begin{cases} f(p_i) - 1 & \text{if } y_i = 1 \\ f(p_i) & \text{if } y_i = 0 \end{cases} \quad (19)$$

The resultant loss value obtained during training is also shown in Fig. 21. The graph shows that, as the learning rate increases, the loss value is getting decreased. This model works under various environments with varying lighting conditions and background. This is because the training has been done on relative positions rather than training using raw pixel values.

Conclusion

Brain-like computing has enormous applications, especially in the healthcare fields, where a huge amount of data has to be processed within a short span of time and enables the system to learn and act accordingly in varying environments. Intelligent assistance makes it very much helpful in healthcare fields that require monitoring people with serious illness. Ambient Assistant Living improves the quality of life for everyone, especially older adults with some diseases. An intelligent system based on posture recognition helps to understand and analyze different kinds of complex data easier using the latest technologies such as deep learning and computer vision. The proposed system implemented a state of the art computing model for the development of public healthcare industry. The real-time pose estimation models have been implemented successfully to classify both normal

and abnormal activities based on binary classification methods. The system also detects other types of activities, such as sitting, standing, walking, and lying on the floor.

In future, brain-like computing model can be utilized for identifying normal and abnormal activities in people having neurodegenerative diseases such as Parkinson's disease. These types of people need careful attention continuously. The poses of these patients are comparatively different from the poses of a normal human being. The model mimics the functionality of the brain to identify various changes happening in the body and act accordingly in real time. Even though the activities performed are the same, the poses may be different. So a careful analysis is needed for activity as well as fall detection. Once the fall is detected, the information can be passed to the caretakers to take immediate action for those who need hospitalization. Combining this method with some wearable devices makes the fall detection both in normal rooms and bathrooms areas. The combined method can be utilized for making a completely automated caretaking system under different environments.

Declarations

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- World Health Organization, World Health Organization. Ageing, and Life course Unit (2008) WHO global report on falls prevention in older age. World Health Organization, Geneva
- Li B, Han C, Bai B (2019) Hybrid approach for human posture recognition using anthropometry and BP neural network based on Kinect V2. *EURASIP J Image Video Process* 2019(1):8
- Patil CM, Ruikar SD (2020) 3D-DWT and CNN based face recognition with feature extraction using depth information and contour map. In: *Techno-societal 2018*. Springer, pp 13–23
- Taramasco C, Rodenas T, Martinez F, Fuentes P, Munoz R, Olivares R, De Albuquerque VH (2018) A novel monitoring system for fall detection in older people. *IEEE Access* 6:43563–43574
- Daher M, Diab A, El Najjar ME, Khalil MA, Charpillat F (2016) Elder tracking and fall detection system using smart tiles. *IEEE Sens J* 17(2):469–479
- Wang H, Zhang D, Wang Y, Ma J, Wang Y, Li S (2016) RT-Fall: a real-time and contactless fall detection system with commodity WiFi devices. *IEEE Trans Mob Comput* 16(2):511–526
- Su Bo Yu, Ho KC, Rantz Marilyn J, Marjorie S (2014) Doppler radar fall activity detection using the wavelet transform. *IEEE Trans Biomed Eng* 62(3):865–875
- Wang YW, Kaishun NLM (2016) Wifall: device-free fall detection by wireless networks. *IEEE Trans Mob Comput* 16(2):581–594
- Selvabala VSN, Ganesh AB (2012) Implementation of wireless sensor network based human fall detection system. *Procedia Eng* 30:767–773
- Andreas E, Matthew B, Lord Stephen R, Janneke A, Redmond Stephen J, Kim D (2016) Wavelet-based sit-to-stand detection and assessment of fall risk in older people using a wearable pendant device. *IEEE Trans Biomed Eng* 64(7):1602–1607
- Forbes G, Massie S, Craw S (2020) Fall prediction using behavioural modelling from sensor data in smart homes. *Artif Intell Rev* 53(2):1071–1091
- Tao X, Sun W, Shaowei L, Ma K, Wang X (2019) The real-time elderly fall posture identifying scheme with wearable sensors. *Int J Distrib Sens Netw* 15(11):1550147719885616
- Caviedes J, Li B, Jammula VC (2020) Wearable sensor array design for spine posture monitoring during exercise incorporating biofeedback. *IEEE Trans Biomed Eng* 67:2828–2838
- Liu J, Shahroudy A, Perez M, Wang G, Duan LY, Kot AC (2019) Ntu rgb+ d 120: a large-scale benchmark for 3d human activity understanding. *IEEE Trans Pattern Anal Mach Intell* 42:2684–2701
- Yang L, Ren Y, Zhang W (2016) 3d depth image analysis for indoor fall detection of elderly people. *Digit Commun Netw* 2(1):24–34
- Pellegrini S, Iocchi L (2008) Human posture tracking and classification through stereo vision and 3d model matching. *EURASIP J Image Video Process* 1–12:2007
- Panahi L, Ghods V (2018) Human fall detection using machine vision techniques on RGB-D images. *Biomed Signal Process Control* 44:146–153
- Li M, Luo X, Yang J, Sun Y (2016) Applying a locally linear embedding algorithm for feature extraction and visualization of MI-EEG. *J Sens* 2016, Hindawi
- Kaur D, Kaur Y (2014) Various image segmentation techniques: a review. *Int J Comput Sci Mob Comput* 3(5):809–814
- Qiu Z, Yao T, Mei T (2017) Learning deep spatio-temporal dependence for semantic video segmentation. *IEEE Trans Multimed* 20(4):939–949
- Zeng D, Chen X, Zhu M, Goesele M, Kuijper A (2019) Background subtraction with real-time semantic segmentation. *IEEE Access* 7:153869–153884
- Hasan S, Samson CS-C (2017) Universal multimode background subtraction. *IEEE Trans Image Process* 26(7):3249–3260
- Lin K-P (2013) A novel evolutionary kernel intuitionistic fuzzy c-means clustering algorithm. *IEEE Trans Fuzzy Syst* 22(5):1074–1087
- Tianming Yu, Yang J, Wei L (2019) Dynamic background subtraction using histograms based on fuzzy c-means clustering and fuzzy nearness degree. *IEEE Access* 7:14671–14679
- Li C, Liu L, Sun X, Zhao J, Yin J (2019) Image segmentation based on fuzzy clustering with cellular automata and features weighting. *EURASIP J Image Video Process* 2019(1):1–11
- Zhao Z-Q, Zheng P, Shou-ao X, Xindong W (2019) Object detection with deep learning: a review. *IEEE Trans Neural Netw Learn Syst* 30(11):3212–3232
- Shih K-H, Chiu C-T, Lin J-A, Bu Y-Y (2019) Real-time object detection with reduced region proposal network via multi-feature concatenation. *IEEE Trans Neural Netw Learn Syst* 31:2164–73
- Suresh D, Priyanka T, Rao EN, Rao KG (2018) Feature extraction in medical images by using deep learning approach. *Int J Pure Appl Math* 120(6):305–312

29. Laulkar CA, Kulkarni PJ (2020) Integrated yolo based object detection for semantic outdoor natural scene classification. In: *Applied computer vision and image processing*. Springer, pp 398–408
30. Redmon J, Farhadi A (2018) Yolo v3: an incremental improvement. arXiv preprint [arXiv:1804.02767](https://arxiv.org/abs/1804.02767)
31. Wang X, Feng SL, Yan WQ (2019) Human gait recognition based on self-adaptive hidden markov model. *IEEE/ACM Trans Comput Biol Bioinform*, 1–1. <https://doi.org/10.1109/TCBB.2019.2951146>
32. Zhong Y, Pei Y, Li P, Guo Y, Ma G, Liu M, Bai W, Wu WH, Zha H (2020) Depth-based 3d face reconstruction and pose estimation using shape-preserving domain adaptation. *IEEE Trans Biom Behav Identity Sci*. IEEE
33. Jun L, Henghui D, Amir S, Ling-Yu D, Xudong J, Gang W, Kot Alex C (2019) Feature boosting network for 3D pose estimation. *IEEE Trans Pattern Anal Mach Intell* 42(2):494–501 (Accessed 06 Nov 2020)
34. Wu Y, Kirillov A, Massa F, Lo W-Y, Girshick R (2019) Detectron2. <https://github.com/facebookresearch/detectron2>
35. Skalski P (2019) Towards datascience, gentle dive to convolutional neural networks. <https://towardsdatascience.com/gentle-dive-into-math-behind-cnn-79a07dd44cf9> (Accessed 13 April 2019)
36. Benjamin C, Taeyong S, Heo Hyun M, Bori K, Hyunggun K, Hwan MJ (2018) A novel detection model and its optimal features to classify falls from low-and high-acceleration activities of daily life using an insole sensor system. *Sensors* 18(4):1227
37. Khanh DQ, Gil SH, Duong PD, Youngjoon C (2019) Wearable sensor based stooped posture estimation in simulated Parkinson's disease gaits. *Sensors* 19(2):223
38. Espinosa R, Ponce H, Gutiérrez S, Martínez-Villaseñor L, Brieva J, Moya-Albor E (2019) A vision-based approach for fall detection using multiple cameras and convolutional neural networks: a case study using the up-fall detection dataset. *Comput Biol Med* 115:103520
39. Feng G, Mai J, Ban Z, Guo X, Wang G (2016) Floor pressure imaging for fall detection with fiber-optic sensors. *IEEE Pervasive Comput* 15(2):40–47
40. Mao A, Ma X, He Y, Luo J (2017) Highly portable, sensor-based system for human fall monitoring. *Sensors* 17(9):2096
41. Peng Y, Peng J, Li J, Yan P, Hu B (2019) Design and development of the fall detection system based on point cloud. *Procedia Comput Sci* 147:271–275
42. Shiba K, Kaburagi T, Kurihara Y (2017) Fall detection utilizing frequency distribution trajectory by microwave doppler sensor. *IEEE Sens J* 17(22):7561–7568
43. Wang F, Gong W, Liu J (2018) On spatial diversity in WiFi-based human activity recognition: a deep learning-based approach. *IEEE Internet Things J* 6(2):2035–2047

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