V10-2 (2023) 80-98
International Journal of Computers and Information
(IJCI)
Available online at *https://ijci.journals.ekb.eg/*

Egyptian Knowledge Bank
بنك المعرفة المصري

Human Fall Detection Using Spatial Temporal Graph Convolutional Networks

Hadir A. Abdo, Khalid M. Amin, Ahmed M. Hamad

Information Technology Department, Menoufia University, Egypt
hadir.atef@ci.menofia.edu.eg, k.amin@ci.menofia.edu.eg, ahmahit@ci.menofia.edu.eg

Abstract

Falls are a serious issue in society and have become a major topic in the healthcare domain. Because of the rapidly increasing number of elderly people, falling can cause serious consequences for the elderly, especially if the fallen person is unable to get up. Early detection of falls and reducing waiting times help save the lives of the elderly. The increasing number of cameras in our daily environment, coupled with the presence of a smart environment, makes the vision-based system the optimal solution for fall detection tasks. A vision-based system using convolutional neural networks (CNN) to detect a fall event in different scenes with different background models is proposed in this paper. For privacy concerns and to avoid complex background problems, skeleton data for the human body was used as an input to the network. A pre-trained spatial temporal graph convolutional network (ST-GCN) model is used for the fall event classification task. ST-GCN classifies the extracted spatial and temporal features from the skeleton data of a detected human as falling or non-falling. To evaluate the proposed system, three public datasets (FDD, URFD, and MCF) that have different environmental issues are used. The experimental results prove the efficiency and robustness of the proposed system in complex situations. The proposed system achieves high performance rates compared to several state-of-the-art systems, with an overall accuracy of 98.6%.

Keywords:Fall detection;Deep Learning;Human Detection;Skeleton Data;Graph Convolution Neural Network (ST_GCN).

*1.*Introduction

According to a UN report released by the World Health Organization (WHO) on ageing and health, the number of elderly people aged over 60 will be doubled by 2050, and their number will exceed the number of teenagers. [1]. As a result, there is a growing worry regarding the elderly health and safety. One of the causes of fatalities and major injuries among the elderly is falling. A prompt response to elderly patients may help to keep them safe from hazardous situations or perhaps prevent their death. As a result, it's crucial to quickly provide medical assistance to senior people after a fall in order to preserve their health. So, technologies for automatically detecting falls have been developed.

The goal of automatic fall detection systems is to identify a fall accident and provide a warning in the event of a threat, assisting the elderly in getting the right medical attention when they need it [2]. Automatic fall detection systems can be classified into wearable, ambient, and vision-based fall detection systems. Elderly persons can be tracked anywhere owing to sensors in wearable gadgets like switches and accelerometers. The elderly is forced to wear these devices frequently, they are easily forgotten, and they require constant recharging. Sensors are built into ambient devices and are positioned everywhere around the monitored object [3]. It could be challenging to adapt ambient gadgets to various living situations.

Vision-based fall detection systems are recommended [4] because of the challenges with wearable and ambient systems that have already been discussed previously as well as the proliferation of cameras in our everyday surroundings. Additionally, the camera offers a wealth of facts about the scene, which contributes to achieving highly accurate results. The vision-based approach also has no sensory side effects on the health of the elderly and has no negative consequences on their ability to lead regular lives.

There are many strategies for detecting falls in the vision-based fall detection approach, which can be divided into two main categories: traditional methods [5] and deep learning methods [6]. In the traditional methods, the moving object is detected, hand-crafted features are collected, and then the classification task is applied. The first step in traditional methods for vision-based fall detection is object detection. There are several techniques for detecting moving objects during the object detection phase, including optical flow, frame difference, and background subtraction methods. The optical flow method can be significantly impacted by environmental noise [7]. The frame difference and background subtraction techniques can also be challenging, since noise and illumination conditions in the surroundings change over time, making it difficult to precisely detect the object [8]. The extraction of the hand-crafted features, which has a high error rate, was the next step in the traditional method techniques. The feature extraction step is also sensitive to challenges such as the occlusion problem, background clutter, or change in camera angle [9]. Then, the classification task utilizes the features captured from the elderly image analysis, such as head movement speed, body length, and so on, to determine the event of a fall. The accuracy of the classification of the elderly's actions as falling or not falling is influenced by the difficulties of moving object detection and the features of extraction steps. The reviews [10] and [11] presented explanations of the fall detection traditional methods and operational flow.

Deep learning methods are developed to address the drawbacks of fall detection traditional methods since deep learning approach excels in the field of image processing and achieves higher accuracy and better performance in the detection of vision falls. Convolution neural networks (CNN) uses a number of object detection techniques for person detection in the human detection phase [12,13]. The object detection network, in short, segments the image into grids, predicts two bounding boxes for each grid cell, and then looks for the best bounding boxes over the entire image. CNN models swiftly and precisely detect moving objects in the image using depth data that was retrieved from the deep convolution layers. It thereby resolves the drawbacks of traditional approaches. The distinctive features are then automatically extracted by CNN model during the training stage. In order to extract depth information from input frames, several convolutions are carried out. The classification process uses the fully connected layers and the pooling layer to classify the fall event based on the features that were extracted during the training phase, thereby reducing calculations and parameters. Based on the aforementioned, deep learning methods outperform the fall detection traditional methods since it has the highest precision and robustness besides resolving the problems that traditional methods had, such as occlusion, background clutter, changes in illumination, and changing camera view angles.

The paper is organized into the following sections: Section 2 reviews the current methods for deep learning-based fall detection systems. Section 3 provides an overview of the proposed method. Section 4 presents the experimental results. Conclusions are presented in section 5.

2. Related Work

Deep neural networks are being used in an increasing range of computer vision applications, including the vision-based fall detection. Using convolutional neural networks increases adaptability and decreases dependency on human resources. It also addresses control of the problems with traditional fall detection techniques. Yong Chen et al. introduced a deep learning-based solution for vision fall detection in [14]. During the object detection phase, pre-trained Mask R-CNN is first utilized to identify moving objects from input frames. The next step is featuring extraction using the convolutional neural network's VGG16 model. Once the characteristics are recovered, the attention-guided bi-directional LSTM model is utilized to pinpoint fall events. The visual attention model concentrates on the key geographic areas of autumn events. The Bi-directional LSTM predicts the classification outcomes of input sequences using both forward and backward temporal information. The VGG model used in the feature extraction phase of the proposed approach is slower than more current models, and the LSTM is time-consuming and prone to overfitting, despite the fact that it can accurately detect falls in movies and outperform state-of-the-art techniques. CNN with two streams: a spatial CNN stream with raw image difference and a temporal CNN stream with optical flow was proposed in [15]. In contrast to conventional two stream action recognition work, Ge, Chenjie et al. use sparse representation with residual-based pooling on the CNN derived features to provide more discriminatory feature codes. To classify the sequential data in video frames, the code vector was created from the long-range dynamic feature representation by recombining codes at segment-levels. The experiments for fall detection have focused on two different video datasets. Despite the system's success when compared to other research, it did not offer any specific information on how to deal with occlusion problems or take pictures from different camera angles.

Nez-Marcos et al. introduced a transfer learning-based fall detection method in [16]. The optical flow images of the input frames were used for the object detection task. Optical flow images depict the motion of two subsequent frames that are separated from one another too briefly to discern a descent. Then, to extract various features, a CNN was fed a stacked collection of optical flow images. As a classifier, a modified VGG-16 network was employed; by stacking several of them, the network can also learn longer time-related features. A fully connected neural network (FC-NN) classifier used these features to determine if a signal indicated a "fall" or "no fall." Despite the fact that they have a strong capacity to provide excellent representational power for motion and optical flow pictures, sequential frame pre-processing puts severe constraints on illumination changes and imposes a large computational overhead.

A fall detection system based on the Kinect sensor is presented in [17]. In order to represent human posture, the skeleton of the human body is extracted using Kinect sensor. But Kinect sensor cannot accurately track every main joint in the human body. Then these features (angle, velocity, and distance) are defined and computed for a detected key joint. Then, the plane's distance from the room floor and its velocity with respect to the head and spine is calculated. The SVM was then used to classify these calculated features. In this method, the authors stated that some fall-like samples, such as lying down or sitting, had to be excluded in order to achieve best results. Due to noise, using the Kinect sensor to detect persons is challenging. In addition, the traditional approach used to compute angles and velocity has a high likelihood of mistakes, making it difficult to precisely detect the fall.

The authors of [18] employ a vision-based fall detection system that acquires the foreground depth data using Microsoft's Kinect V2. To reduce the noise and extract skeletal data in a thin shape, the dilation and erosion operations are used. The likelihood of a wrong computation is decreased because the skeletal information is now clear after the thinning. Seven joints above the waist are used to determine the fall since those joints will change significantly in the case of a fall. This streamlined joint-point set allows for the reduction of the training parameters. Finally, these extracted features are provided to CNN in order to categories the fall event. The high price of the Kinect camera making it out of the reach for the majority of people. Additionally, there are a lot of processes involved in getting a well-suited input for the model, which adds to the time and error risk.

Weiming, et al. used the open-pose algorithm to acquire human body skeletal data [19]. Then, the key aspects that control how the human body moves are calculated, including the angle between the human centerline and the ground, the width-to-height ratio of the external bounding box, and the velocity of fall at the hip joint's center. The task of classifying falls is then accomplished by examining the following features: (i) determining whether or not the hip joint's center of descent velocity is greater than the critical velocity, (ii) determining whether or not the human body's width to height ratio is greater than 1, and (iii) determining whether or not the angle between the human centerline and the ground is less than 45 degrees. The used dataset did not cover people's behaviour and actions when there was a partial occlusion, and the action was detected from the side without considering the other directions, even though the technique used is likewise simple to apply and inexpensive. Additionally, hand measurements of the extracted features take a lot of time and increase the rate of classification mistakes.

Recent deep learning approaches that are based on graphical data gained increased interest for fall detection because of its great expressive power of graphics [20]. Graphical data in the case of human activities leverages the natural connections of human body. These natural connections that are defined as edges and joints are used for recognizing human actions. These graphical data are robust to illumination change, scene variation, and are easy to obtain with high accuracy. In recent years, there has been a lot of skeleton-based action recognition research released recently.

To get over the aforementioned challenges discussed in this section, the proposed method presented in this paper leverages deep learning based on skeletal data. The proposed method detects the fall event through a set of consecutive CNN models. The Graph Convolutional Network (GCN), a very effective neural network architecture for graph machine learning, was utilized based on skeletal data. The proposed method first uses the YOLO model for person detection. YOLO has the highest accuracy and operates in real time when compared to other object detection models. The proposed method therefore relies on YOLO to precisely locate the human over a set of frames. Then, the skeletal information for the detected human is obtained using open pose technique The spatial-temporal graph convolution network (ST-GCN) classifies the person fall through offering a high level of precision and privacy protection. In order to determine if human motion is falling or not, an ST-GCN network uses temporal and spatial data for classification tasks. The experimental results show that the suggested methodology is effective, outperforming state-of-the-art approaches by up to 98.6% accuracy, and are used to assess how well the proposed framework works with the URFD, FDD, and MCF datasets.

3. **Proposed Work**

The proposed Graph Convolution Neural Network (GCNN) based fall detection system workflow is shown in Fig.1. The proposed fall detection system has three phases. The first phase attempts to detect moving humans in video frames using the YOLO model. The second phase is performed using an Open Pose model that extracts the human body skeleton data key points from RGB frames resulting from the YOLO algorithm. The skeleton and joint trajectories of human bodies are robust to scene variations and illumination changes. The output skeleton is a graphical representation that consists of nodes and vertices. The node represents each main part of human body and vertices are used to establish relationships between pairs of nodes. Finally, the third phase uses a Temporal Graph Convolutional Network (ST-GCN) network model to analyse human skeleton joints to classify human activity as falling or not falling. The YOLO model was the first tool used to detect moving humans in video frames.

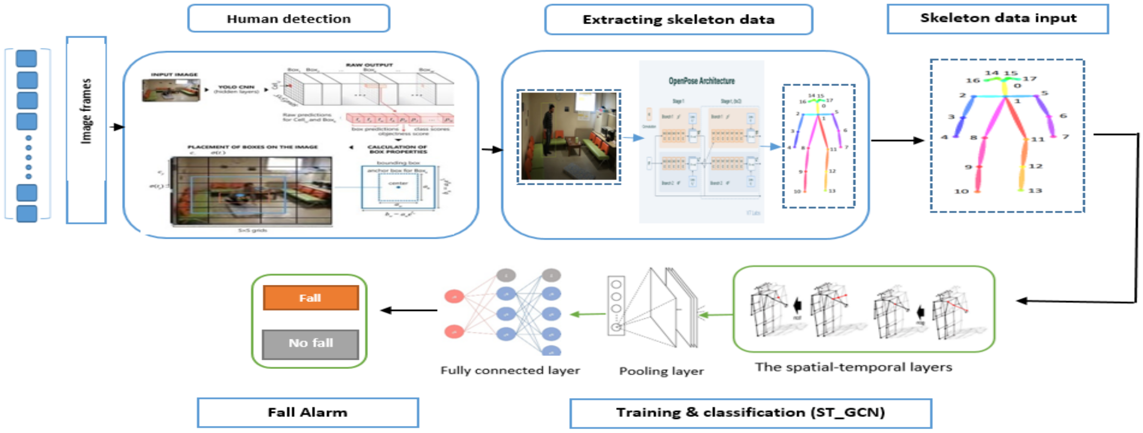


Fig. 1 The workflow of the proposed human fall detection system.

3.1 YOLO Algorithm

The YOLO algorithm is a real-time object detection technique that recognizes moving objects in image frames. YOLO is distinguished from other CNN models for object detection by its speed and accuracy. YOLO employs three consecutive techniques: residual blocks, bounding box regression, and inter-section over union (IOU). First, in the residual block, YOLO divides an input image into grids with a dimension of S*S. Each grid cell is responsible for detecting objects (persons) that appear inside this grid [21]. A grid cell predicts one object for each box confidence score. The number of predicted objects and bounding boxes in a grid cell is directly proportional. Second, YOLO employs bounding box regression, and each bounding box can be identified by four properties: bounding box width (bw), bounding box height (bh), bounding box center (bx, by), and class score, which is identified by the letter c and refers to any class the object belongs to (e.g., humans, cars, animals, etc.).Third, the intersection over union (IOU) is to be utilized to assess how much the predicted bounding box resembles the actual box. In addition, IOU removes bounding boxes that aren't necessary or don't match the properties of the actual objects. If the IOU is equal to 1, the predicted bounding box is said to be identical to the real box. Unique bounding boxes for each human appearing throughout the time frame of the frame sequence make up the final detection. The YOLO algorithm model architecture is shown in Fig.2.

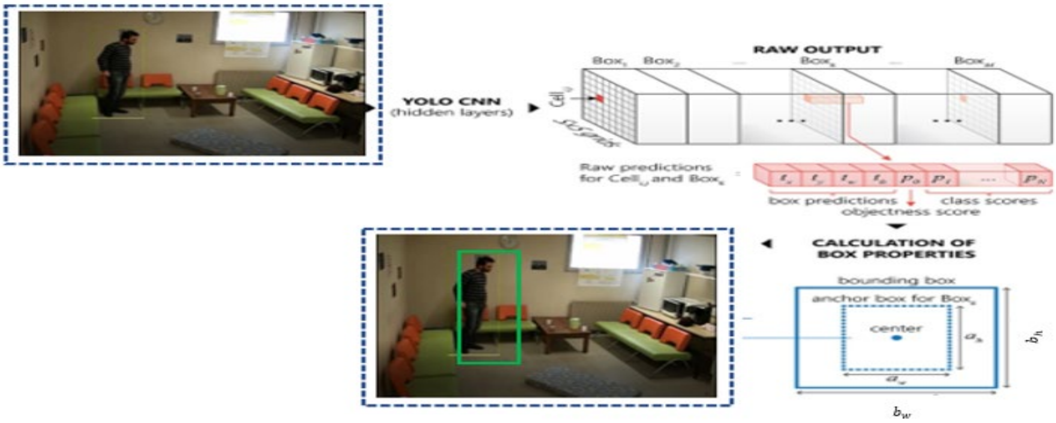

RAW OUTPUT
Box$_1$ Box$_2$ ... Box$_b$ ... Box$_M$
YOLO CNN
(hidden layers)
Cell$_{ij}$
S×S grid
Raw predictions for Cell$_{ij}$ and Box$_b$ =
t_x t_y t_w t_h p_0 p_1 ... p_N
box predictions
objectness score
class scores
CALCULATION OF BOX PROPERTIES
bounding box
anchor box for Box$_b$
center
a_h
a_w
b_h
b_w

Fig. 2 The YOLO algorithm architecture for human detection.

3.2 Open Pose Algorithm

The proposed method utilizes the Open Pose model to extract the skeletal data of the detected humans. Open Pose is a real-time multi-person CNN architecture to identify human body, hand, facial, and foot key-points. The output of the YOLO model is fed into the Open Pose model to obtain skeletal information on human body nodes for each frame [22]. As shown in Fig. 3, the input frames are processed by Open Pose to produce nodes for each major body part. Using an open pose, 18 human nodes have been found in the human body. The output of the open-pose algorithm is represented as an array of 2D coordinates (X, Y, and C). In a 2D array, X and Y represent the node location, and C represents the confidence score for each node, as each node is one of 18 nodes in the human body (left ear, right ear, left shoulder, right shoulder, left hip, etc.). As shown in Fig. 4, the skeleton frame is stored as an array of 18 tuples, with (X, Y, C) for each node in each tuple. The array of tuples representing human node is used as the input for the spatial-temporal graph convolutional network (ST-GCN), which is used to classify falls in the proposed method.

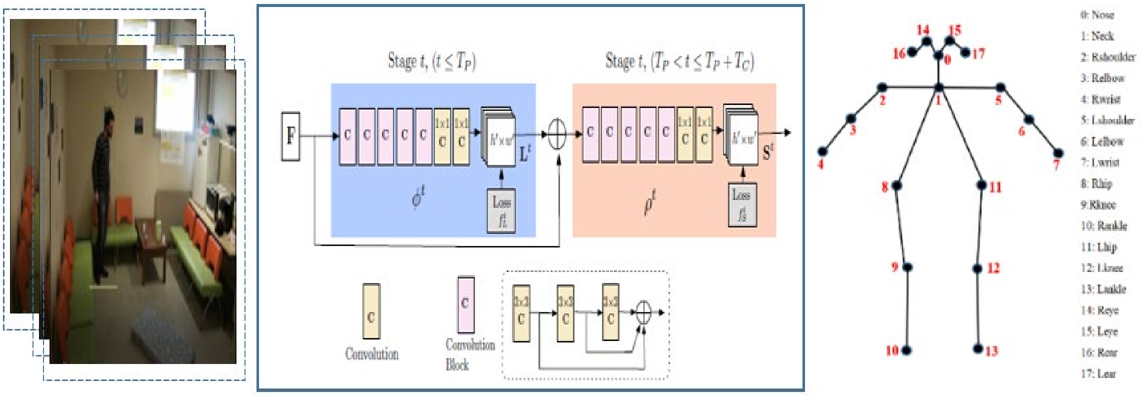


Fig. 3 The architecture of the Open Pose algorithm.

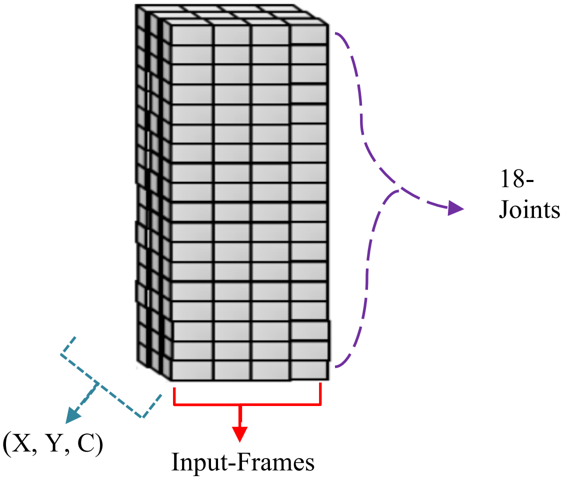


Fig. 4 The video clip representations of the Open Pose algorithm outputs.

3.3 Spatial Temporal Graph Convolution Network

Due to the size limitation of the fall detection datasets, a limited number of deep learning architectures are used to detect human falls. To get around this constraint, transfer learning is a deep learning technique that is commonly applied in order to allow models that have already been trained to serve as the foundation for models on related tasks. In transfer learning, the learned features are first applied to a base network trained on a large dataset for a specific task. Then, these learned features are transferred to a second target network that is trained on a target dataset and related task.

Because the fall detection dataset sizes are not large enough, the proposed method uses transfer learning to transfer the learned features for the human action recognition task, which is carried out through the Spatial-Temporal Graph Convolutional Network (ST-GCN), to specific features for the human fall event. ST-GCN is a deep learning approach that uses the skeletal data to categorize human motions [23]. The spatiotemporal pattern of skeletal joints can be captured by the ST-GCN. Each joint point is identified as a node in the skeleton graph, while the connections between nodes are identified as edges. The connections between nodes function as a spatial pattern in one frame. The similar nodes across successive frames are connected as a temporal pattern in neighbouring frames. Interbody and intra-body edges are the two different categories of edges. The interbody edges link each node to itself over the course of several frames, and they serve as a representation of the internal connections shown in human bodies. A spatial and temporal graph convolution layer is represented by each ST GCN unit in the ST GCN network. The nodes and joints present the input data graphically. The spatial convolutional layer receives the input data and extracts the spatial features for each joint. The graph convolutional neural (GCN) layer is an extremely potent serves as the foundation for the spatial convolutional layer. The output of the spatial convolutional layers is then fed into the temporal convolutional layer (TCN), which extracts the temporal features of the same joint over a series of frames. The extracted features are then assigned to the pooling and fully connected layers for classification.Fig.5 illustrate the architecture of the Spatial-Temporal Graph Convolutional Network.

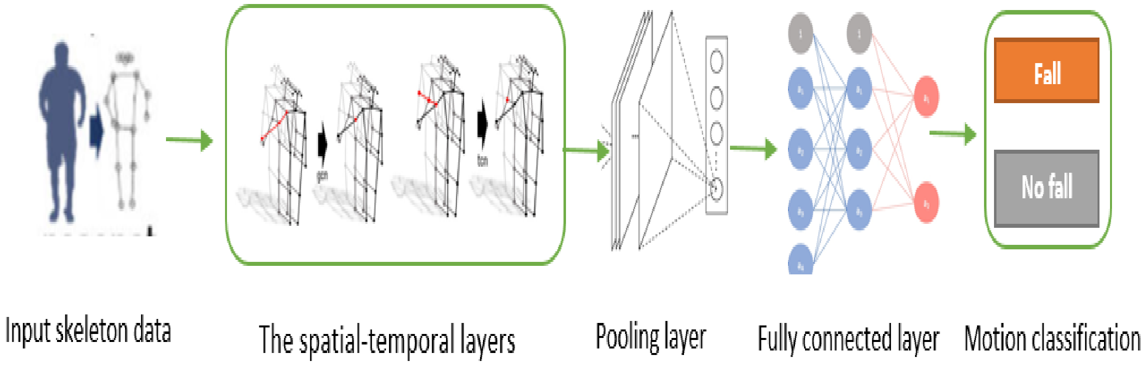


Fig. 5 The Spatial-Temporal Graph Convolutional Network architecture.

For node labelling, the ST-GCN employs spatial configuration partitioning during the labelling process. As shown in Fig.6. (A), the centripetal groups (blue nodes) have shorter distances to the centre of gravity than the root node, while the centrifugal groups (yellow nodes) have longer distances. The nodes in the spatial configuration partitioning are labelled according to their distances from the centre of gravity (black cross). Following the labelling process, the network's spatiotemporal layers examine the position and relationships between each identified node and the nodes in the same frame as well as the same node and the nodes in subsequent frames. Therefore, by using the spatial-temporal layers to extract structural information from the human body, the spatial-temporal information is obtained. The subsequent pooling and fully connected layers, which have high relational modelling skills, then categorize human motion as falling or not falling by exploiting the temporal and spatial pattern between joints. On the right side of the image, as shown in Fig.6. (B), we also exhibit the characteristics of an ST-GCN convolutional unit. The ST-GCN has the benefit of providing a better depiction of human activity because of its strong relationship modelling capabilities and ability to interact with data in their native form.

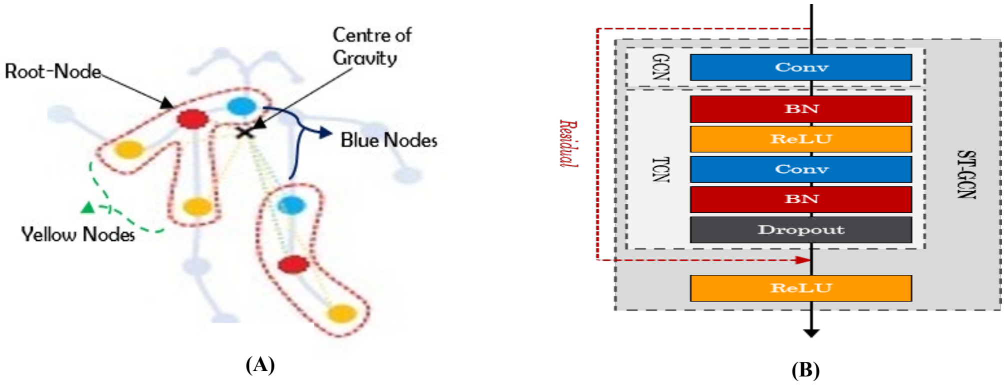


Fig. 6. (A) The spatial configuration partitioning strategy. (B) The spatial-temporal graph convolution network unit.

Fig.7. illustrates the hierarchy of stacked spatial-temporal blocks that the ST-GCN consists of. Each of these blocks is fundamentally made up of a spatial convolution (GCN) and a temporal convolution (TCN). There are nine levels of ST GCN blocks in the ST GCN. The output channels are 64 for the first three levels, 128 for the next three levels, and 256 for the final three levels. The residuals network mechanism and dropout with a 0.5 drop rate were used for each ST GCN unit to solve the overfitting issue. The SoftMax classifier was then used to classify the generated feature vector.

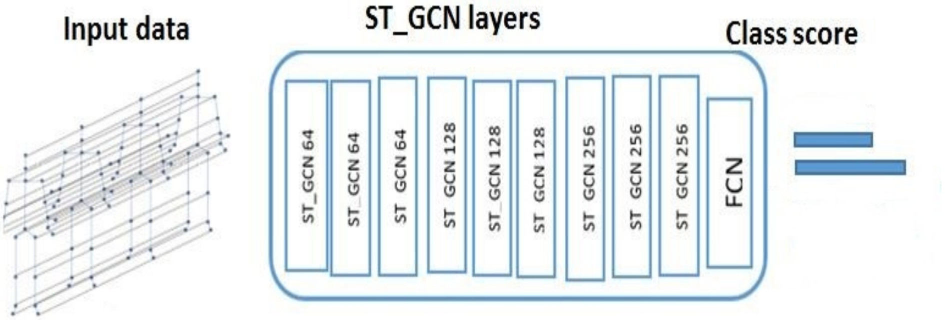


Fig. 7.The Spatial-Temporal Graph Convolutional Network hierarchy.

4. Results and Discussion

4.1 Dataset and Data Pre-processing

The proposed method is tested and evaluated using three datasets: the FDD fall detection dataset, the UR fall detection dataset (URFD), and the multiple cameras fall dataset (MCF). The dataset video sequences include challenges such as varying lighting conditions, occlusions, different camera view angle, multiple persons in the same scene and textured background.

The FDD fall detection dataset was recorded in the Laboratory Electronic Information and Image [24]. The dataset contains 191 videos that were used for experimentation. These videos include human falls and other human activities that occur in daily life. The frame rate is 25 frames per second, and the resolution is 320 x 240 pixels. FDD videos were recorded in different locations (such as "house," "coffee room," "office," and "lecture room"), allowing for a variety of evaluation techniques. Also, videos commonly have a textured background and shadows.

The URFD fall detection dataset contains 70 videos (30 videos for falls and 40 videos for daily human activities). Fall events are recorded with two cameras, including frontal and overhead video sequences, one situated on the floor at a height of 1 metre and the other at a height of 3 metres on the ceiling. Five individuals performed two different types of falls in the dataset: one while standing and one while seated in a chair. The dataset video resolution is 640 x 240 with a frame rate of 30 fps. The dataset included different scenes with illumination changes and various camera angles. In addition, some videos include multiple people, which may lead to a partial occlusion problem [25].

The multiple cameras fall dataset (MCF) is one of the largest datasets that is regularly used to detect human falls. This dataset contains 24 situations, each of which represents one of nine various actions, such as walking, lying down, crouching, falling, etc. Eight separate cameras were used to capture each activity. Each video has a frame size of 720 x 480 and a frame rate of 30 fps. Additionally, the background of several shots is congested and grainy. The 14 videos from the MCF dataset that were recorded at camera position 2 are described in Table 1 below. Table 1 includes information such as the number of videos, total frames in each video, frames with falls, and frames without falls. [26]

Table 1. *Details of multiple cameras fall (MCF) dataset.*

Video	Total Frames	Fall	No fall	Video	Total Frames	Fall	No fall
Video 1	1114	314	1079	Video 8	700	240	460
Video 2	756	331	425	Video 9	905	360	545
Video 3	883	172	611	Video10	813	230	583
Video 4	1033	409	624	Video 11	1486	608	878
Video 5	600	266	334	Video 12	1041	420	621
Video 6	1203	513	960	Video 13	1240	360	880
Video 7	912	290	622	Video 14	970	385	585

The pre-processing implementation was conducted online on Google Colab (Collaboratory virtual machine) with free GPU and memory resources on the cloud. The used dataset was uploaded to Google Drive, and YOLO v5 was cloned from the GitHub repository to the drive. After mounting Google Drive, the dataset could be used, and the dependency packages and libraries were installed. Then the file named 'coco128.yaml' was customized according to the used datasets and annotations. Finally, training and testing were conducted for the detection of humans within the input frames. For extracting the skeleton data, Open Pose was cloned from the GitHub repository to the Google Colab GPU runtime. CMake with CUDA10 and other prerequisites needed for running have been installed. Then the output frames from YOLO were assigned to Open Pose for keypoint detection. The output of OpenPose is the input frames with identified keypoints for the 18 main body parts of the human body detected and saved as AVI files, and the identified keypoints of the detected human body were saved in PKL files. The output frames are resized to a resolution of 340*256. Finally, the number of output frames multiplied by the 18 joint points of training data is the input for ST-GCN for human fall detection.

4.2 Training and Implementation

Training and testing of the proposed method are implemented using the PyTorch deep learning framework. Training is conducted online on Google Colab (Collaboratory virtual machine) with free GPU and memory resources on the cloud. The testing and evaluation values of the proposed method have been derived from 50 training epochs of the trained model.

Transfer learning from the pre-trained ST-GCN onto the Kinetics-skeleton dataset [27] is utilized to reduce overfitting and enhance model performance. Based on the pre-trained model, initialized weights of the ST-GCN feature extractor are used, and the first nine convolution layers are frozen. We changed the output layer dimensionality to fit the fall detection problem. ST-GCN's output layer is modified to have two classes (fall and no-fall). To fit the pretrained model with the fall detection dataset, the ST GCN network is trained for 50 epochs using the FDD dataset. In the proposed approach, fall event classification is performed by the ST-GCN network using Open-Pose model posture estimation features. The model learned the important features of human motion for a fall or no-fall event during the training phase [28].

There are 25,171 video clips for various motions in the FDD dataset. The training set is made up of 15,840 clips; 3,960 are used in the validation set, and 5,371 are used in the test set. The gradient descent (Ad-Delta) optimizer utilized the following hyperparameters to develop the model: 0.001 learning rate and 0.0001 weight decay A training set made up of eighty percent of the data and a test set made up of twenty percent of the data are randomly partitioned. the loss function that was determined using training and validation data for each learning period. A dropout procedure with a probability of 0.5 was used after each ST GCN unit. The training method consists of 50 epochs. ST-GCN evaluates the model and ensures that it fits the data during the learning phase by using the cross-entropy loss function. The cross-entropy loss, which is the likelihood of the predicted class in relation to the actual class [29], is based on how closely the output values resemble the real value. It illustrates the predictions' convergence with the actual class. For the classification task of fall detection, the cross-entropy loss function is calculated for two classes (classes 0 and 1) as it indicates fall or non-fall and is calculated as follows:

$$L_{CE} = -\sum_{i=1}^{2} T_i \, log(P_i) \qquad (1)$$

In Equation (1), L_{CE} is the cross-entropy loss function. Since i is a class value, T_i represents the actual value for i classes, and P_i is the predicted value of ST GCN learning output (ArgMax values). Assuming that a fall represents itself with a value of 1 and that a non-fall represents itself with a value of 0, the cross-entropy loss is the sum of the values of the actual fall with regard to the output generated from the model (ArgMax values) [30].

4.3 Quantitative Evaluation

Three datasets for fall detection (FDD, URFD, and MCF) are utilized to verify the effectiveness of the suggested fall detection system. All of the dataset’s videos depict actual scenes with various daily activities, including falling. Common criteria for quantitative evaluation include sensitivities, specificity, and accuracy. Sensitivity is also called the "true positive rate" or "recall value," and specificity is the "true negative rate" [31]. Accuracy is the percentage of correctly classifying falls, and it is determined using equation (2). A precision of indicates that each object belongs to the true positive class and is actually classified as such, which is computed using equation (3). The number of non-fall occurrences that are correctly identified as non-fall events is known as specificity (or true negative rate), and it is computed using equation (4). The number of falls that are accurately detected is known as sensitivity, often referred to as recall or true positive rate, and it is computed using equation (5). The weighted average of precision and recall is represented by the F1-score (6).

$$Accuracy = (TP + TN)/(TP + TF + TN + FP) \qquad (2)$$

$$Precision = TP/(TP + FP) \qquad (3)$$

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

$$Senstivity/recall = TP/(FN + TP) \qquad (5)$$

$$F1\_Score \quad = ((TP)) / ((TP+1/2\ (FN+FP))\ ) \qquad (6)$$

For the fall detection problem, the confusion matrix comprises true positive, true negative, false positive, and false negative [32]. According to the confusion matrix shown in Fig. 8, TP represents actions labelled and predicated as "fall" while FP represents actions labelled "no-fall" but predicated as "fall.". The actions labelled "no-fall" and predicated as "false negatives" are expressed by TN, where FN represents the actions labelled "fall" and predicated as "no-fall." The accuracy, sensitivity, specificity, precision, and F1-score metrics results of the proposed framework based on three datasets are shown in Table 2. An overall accuracy of 98.6% was achieved. To ensure the robustness of the proposed method, several comparisons with state-of-the-art methods based on the same evaluation datasets have been conducted.

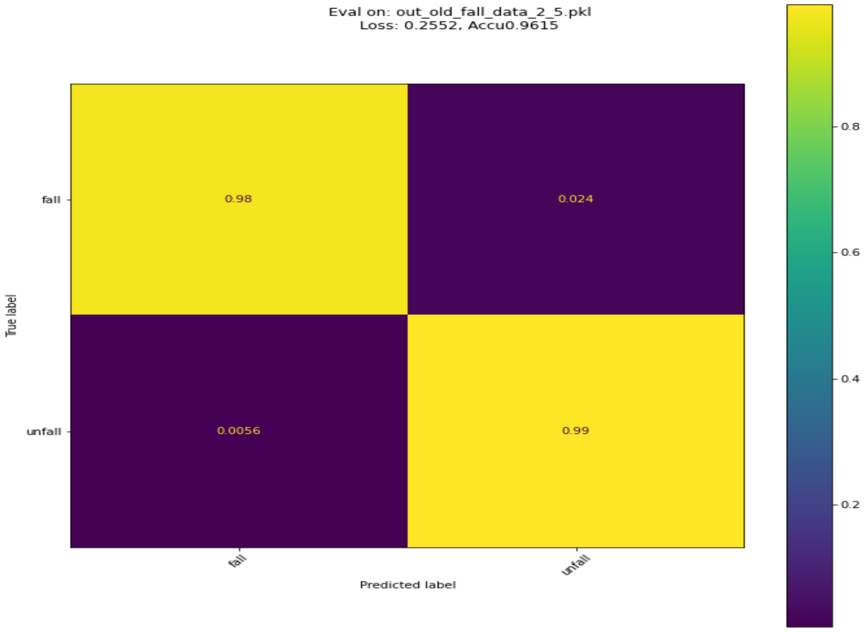


Fig. 8. The confusion matrix obtained for the FDD dataset.

Table 2. Results of the proposed method on the three datasets (FDD, URFD, and MCF)

Dataset	Accuracy	Sensitivity	Specificity	Precision	F1 Score
FDD	98.6%	99.34%	97.7%	97.6%	98.6%
URFD	98.2%	99.4%	97.6%	97.6%	97.4%
MCF	97.5%	98.4%	96%	96.3%	97.0%

In comparison to the URFD and MCF datasets, the FDD dataset has been used more frequently in fall detection experiments. Table 3 compares the outcomes of the proposed method based on the FDD dataset with those of the fall detection methods developed by Alaoui et al. (2019) [33], F. Harrou et al. (2019) [34], and M. M. Hasan et al. (2020) [35]. In terms of accuracy, FI-score, and precision, the proposed approach outperformed other compared methods and achieved the best results (99.34% of the falling scenes correctly and 97.7% of the non-falling sequences correctly). For human detection purposes, the proposed method uses a YOLO algorithm, which works in real time with high accuracy and speed. It also depends mainly on the structural data of the human body; as compared to RGB or grayscale frames, it more clearly shows the various human postures. It unquestionably protects privacy as well, which helps to avoid issues with lighting, shifting camera angles, and different human postures. Additionally, it uses a spatial-temporal neural network (ST GCN) to classify the skeleton input data, which classifies the input frames in both spatial and temporal terms. The ST_GCN proved its efficiency in recognizing the fall and avoiding any false alarm in case of other human activities that have the same characteristics.

Table 3. *Comparison between the proposed method and other State-of-the-art methods tested on the FFD dataset.*

Methods	Alaoui et al [33]	F. Harrou et al [34]	M. M. Hasan et al [35]	The proposed method
Sensitivity	95%	٩٩,٩ %	99%	99.34%
Precision	--	٩٤,٠%	--	97.6%
F1_score	--	٩٦,٦٦%	94.8%	98.5%
Accuracy	97.5 %	٩٦,٦٦%	93.8%	98.6%

Alaoui et al. [33] employed human skeleton data to propose a vision fall detection system using RGB images. In the first step, a 15 main key points of human body were detected by using a 2D skeleton model based VGG-19. After that, the distances and angles between each two pairs of sequential points was calculated. Then, the principal component analysis (PCA) was applied to unify dimension of features. Finally, SVM, KNN, Decision Tree and Random Forest were used as a classifier to classify fall and non-fall videos based generated features (distances and angles). The method was evaluated while the camera in the same place for all time, and according to the authors, it does not perform well when there are variations in lighting.

F. Harrou et al. [34] identified falls using SVM classifier with a generalized likelihood ratio (GLR). The SVM classifier was used to identify falls from other common activities because it can handle both linear and nonlinear information by utilizing a nonlinear kernel. In contrast, using the GLR alone makes it impossible to differentiate between true fall and behaviours that resemble it. This method's accuracy, which is 96.66%, is lower than that of the Alaoui et al. [33] method and the proposed method.

M. M. Hasan et al. used a recurrent neural network (RNN) with LSTM in [35] to detect falls. They constructed a recurrent neural network (RNN) with an LSTM to model the temporal dynamics of a fallen person's 2D pose information. Human 2D pose information, which has been shown to be effective in analyzing fall events as it ignores people's body appearance and environmental information while capturing true motion information, simplifies and speeds up the proposed method. This method's accuracy, which is 93.8%, has the least accuracy compared with other methods.

Table 4. *Comparison between the proposed method and other state-of-the-art methods tested on the URFD dataset.*

Methods	Adrián et al [36]	Yong -Chen at al [37]	Wang et al [38]	The proposed method
Sensitivity	99%	91.7%	97.76%	9٩.4%
Precision	--	100%	97.78%	9٧.٦%
F1_score	94%	94.8%	97.76%	97.4%
Accuracy	95%	96.7%	97.33%	9٨.٢%

Table 4 compares the results of the proposed method based on the URFD dataset with the results of the fall detection method developed by Adrián Nez-Marcos [36], Yong-Chen [37], and B. Wang et al [38]. In terms of accuracy, FI-score, and precision, the proposed method achieved the highest results among the comparable methods using the URFD dataset. The proposed method proved its ability to recognize human falls with potential changes. In order to avoid any false alarms in the case of stillness and overlap between the background and detected humans, the proposed method employs the YOLO algorithm, which works in real time with high accuracy and speed to recognize humans in the video frames. Since the skeletal data is more accurate than RGB and grayscale frames and respects human privacy concerns, it serves as the main basis for the proposed method for avoiding partial occlusion and the varied camera angles. In addition, the proposed method uses a spatial temporal graph neural network (ST_GCN) which classifies the input frames in both spatial and temporal terms. The ST_GCN proved its efficiency in recognizing the fall and avoiding any false alarm in case of partial occlusion and different human posture.

Optical images were used by Adrián Nez-Marcos et al. in [36] to categorize human falls since optical images offered more information about movement. Using optical pictures as inputs, the convolutional network categorizes human movements, specifically identifying falls and non-falls. The method used transfer learning of VGG network, which was trained on ImageNet. While optical flow was successful at capturing movement in video frames, it also had drawbacks related to lighting changes and the laborious computing required to prepare successive frames. Adrián Nez-Marcos et al.'s [36] method had the lowest accuracy of the methods tested.

LSTM and VGG CNN models were employed by Yong-Chen et al. in [37] for the problem of classifying human falls. In order to detect humans in the input frames, they first employed a region-based convolutional neural network (R-CNN). The properties that set apart human falls from other human movement characteristics were discovered during the feature extraction stage using a pretrained model (VGG). For the final fall classification, a guided bidirectional long short-term memory (LSTM) model was used. The model performed well with background texture however it performed poorly with occlusion items and several persons in the same scenario. Yong-Chen et al. method was the third best accuracy at 96.7%.

Multilayer perceptron and random forest were employed by Wang et al. in [38] for the purpose of classifying human falls. The YOLO model was first used for human detection to separate moving objects from the background. The human body's position is then extracted using the OPENPOSE approach, and both static and dynamic human body properties—like centroid speed and upper limb velocity—are extracted using a dual channel sliding window model (human external ellipse). A multilayer perceptron and random forest are used to classify human movements, notably falls, in accordance with the retrieved features. According to the experimental findings, the adopted strategy generated improved accuracy and revealed the limitations of the traditional method's fall detection. Wang et al method was the second-best accuracy at 97.3%.

The MCF dataset used eight separate cameras that were used to capture human activity. Due to the several camera perspectives that result in occlusions and considerable changes in the spatial positions, scale, and orientations of the fall events, the MCF dataset differs from single camera-based fall detection datasets and is regarded as being more difficult. Table 5 compares the results of the proposed framework based on the MCF dataset with the results of the fall detection methods developed by Qing Zhen et al. [39], Alexy et al. [40], and Nazar Mamchur et al. [41]. The proposed method`s accuracy result demonstrates its effectiveness in overcoming partial occlusion and numerous variations in human spatial positions, scale, and orientation caused by various cameras. Among comparable methodologies, the proposed method had the highest accuracy. Since the proposed approach employs a real-time object detection algorithm (YOLO), which has a detection rate of up to 99.8% with various illumination changes and varied camera angles, as well as using Open Pose for extracting the human skeletal data, which preserves human privacy and avoids any misclassification in partial occlusion and in the situation of multiple persons, Additionally, the proposed method makes use of a spatial-temporal graph neural network to analyze the input frame both spatially and temporally, producing reliable results for classifying human falls while avoiding any misclassification or false alarms for human posture with similar characteristics and in case of multiple persons in the same scene.

Table 5. *Comparison between the proposed method and other state-of-the-art methods tested on the MCF dataset.*

Methods	Qing et al [39]	Alexy et al [40]	Nazar et al [41]	The proposed method
Sensitivity	٩٩,٩ %	92.0%	--	98.43%
Precision	٩٤,٠%	88.5%	90.0%	96.3%
F1_score	٩٦,٦٦%	91.3%	90.0 %	97.0%
Accuracy	٩٦,٦٦%	89.1%	91.07%	97.5%

To differentiate between the human fall event and daily activities life, Qing-Zhen et al. [39] applied the transfer learning technique of the Inception-ResNet-v2 model. The primary joints of the human body were obtained using the OPENPOSE. The human bone map that was produced by OPENPOSE was then saved as a data set. Then, transfer learning is employed to train the data set to generate a new model. The updated model is then utilized to forecast the fall by contrasting the current input frames with the previously saved bone map. The experimental results will be significantly impacted by illumination changes and obscured objects in the photos, and thus cannot be used in real-time, according to an analysis of these model results. Qing-Zhen et al. [39] method reported an accuracy of 96.66%.

Alexy et al. in [40] proposed VGG-16 convolutional neural network to identify human falls. A dense optical-flow TVL1 approach was used to create a pair of optical flow images. In order to categorize human activities, VGG-16 was fed by the optical flow images produced by dense optical flow TVL1. CNN output is eventually subjected to a temporal filter and prediction threshold in order to reach a final decision on the human fall event. This method usually causes false alarms if someone sits down for an extended amount of time, gets up after falling, or bends down to pick something up from the ground. The accuracy of the trial's findings is 89.6 %.
Recurrent neural networks (RNN) were used by Nazar Mamchur et al. [41] to construct a vision-based fall detection system. The model extracted 17 joints on the human body using the pre-trained TensorFlow Lite CNN real-time Pose Net model. The input key-point sequences were then examined using a recurrent neural network (RNN). By removing the human key point locations and examining the joint point location changes, the algorithm classifies the fall event. This approach provides no information concerning an occlusion issue or multiple people in the same scene, and the method reported an accuracy of 91.07%.

4.4 Qualitative Evaluation

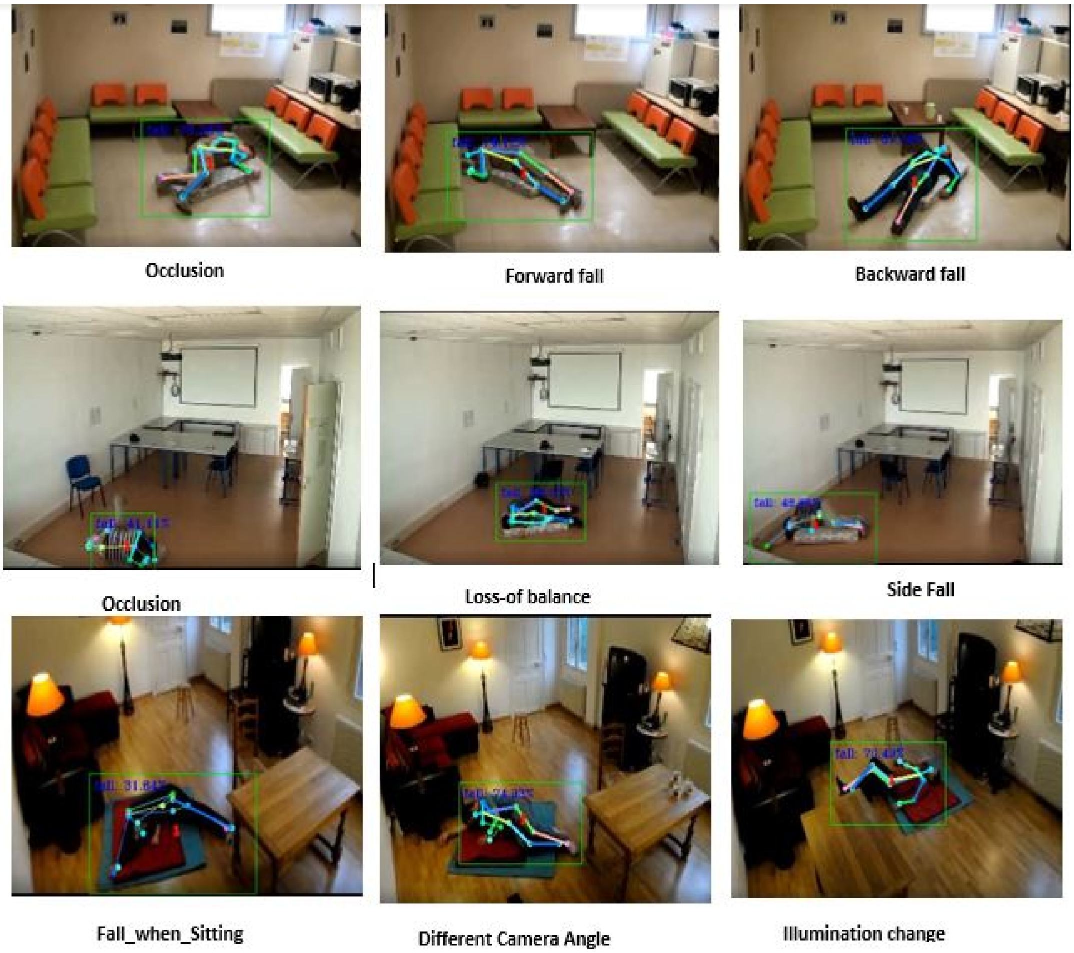
To demonstrate the effectiveness and superiority of the proposed method, qualitative results based on the UFRD, MCF, and FDD datasets were provided. The experimental results of the proposed method on three fall detection datasets are displayed in Figures 9,10 and 11.

In comparison to CNN-based fall detection techniques proposed by Alaoui et al. [33], F. Harrou et al. [34], and M. M. Hasan et al. [35] for the FDD dataset. Also, the earlier results that are being compared were based on CNN approaches, they appear even though poor results in cases of partial occlusion problems as well as false alarms in situations involving various camera angles and human fall postures. Whilst the visual results of the proposed method utilizing the FDD dataset, shown in Fig.9., reported that the proposed method can accurately identify human falls in a number of different situations, The visual results of the proposed method in the case of various partial occlusions and backward and forward falls are shown in the first row. The visual results in the cases of a side fall and a sudden fall are shown in the second row. The third row shows the visual results in the case of changing the camera angle and lighting levels.

Regarding the URFD dataset. The proposed method is compared to CNN-based fall detection techniques proposed by B. Wang et al. [38], Yong-Chen , and Adrián Nez-Marcos[36, 37].The earlier results that are being compared which based on CNN approaches produce poor results and they have no information in cases of multiple persons in the same and some of them skip some daily activity for human from testing results. They also provided no information when there were multiple people involved. While the visual results, as shown in

Fig.10, showed that the proposed approach is capable of correctly identifying human falls in a wide range of situations. The visual results of the proposed approach for various partial occlusions and backward and forward falls are shown in the top row. The visual results of the proposed approach for multiple people in the same picture, side falls, and sudden falls are shown in the second row. The third row shows the visual results of the suggested approach when the lighting and camera angle are changed.

The proposed method is compared to the results of the fall detection techniques proposed by Qing Zhen et al. [39], Alexy et al. [40], and Nazar Mamchur et al. [41] on the MCF dataset. The earlier results that are being compared were based on CNN, and some of them also used skeletal data. The results of the compared methods seem unstable and produce poor results in cases of partial occlusion and multiple people in the same scene. While the visual results of the proposed method showed that the proposed method is able to accurately distinguish human falls in a wide range of situations. Fig. 11 shows the visual results of the proposed method based MCF dataset; the top row displays the visual outcomes of the proposed method for various partial occlusions and backward and forward falls. The second row shows the visual outcomes of the proposed method for multiple people in the same scene, side falls, and sudden falls. The third row shows the method's visual outcomes under different lighting conditions.



Occlusion

Forward fall

Backward fall

Occlusion

Loss-of balance

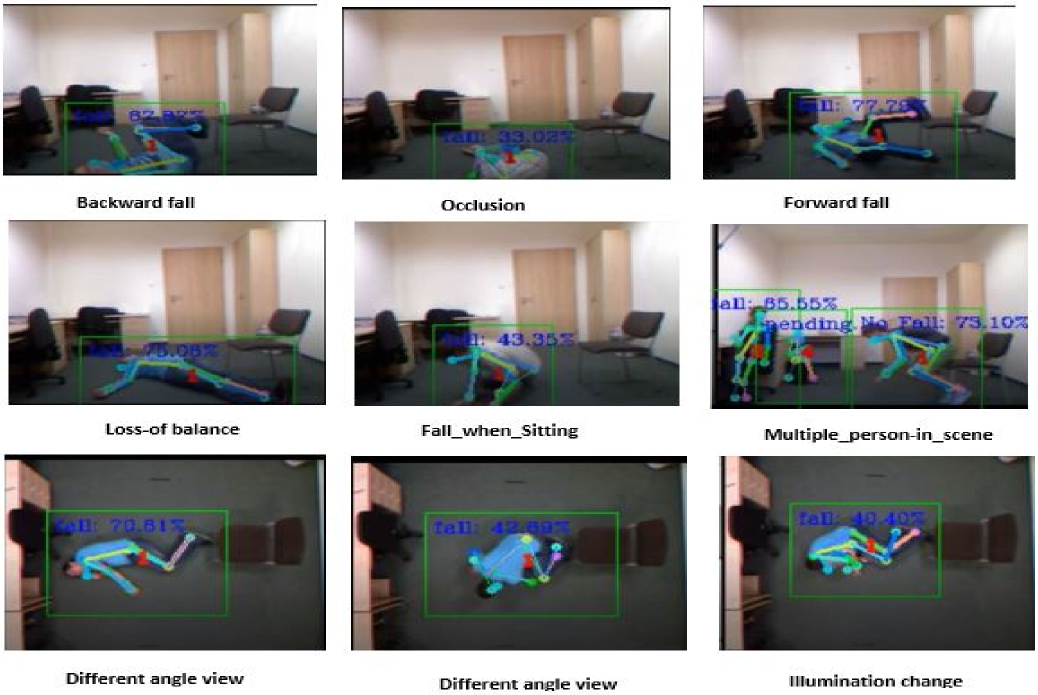
Side Fall

Fall_when_Sitting

Different Camera Angle

Illumination change

Fig. 9. The experimental results using the FFD dataset.


Backward fall

Occlusion

Forward fall

Loss-of balance

Fall_when_Sitting

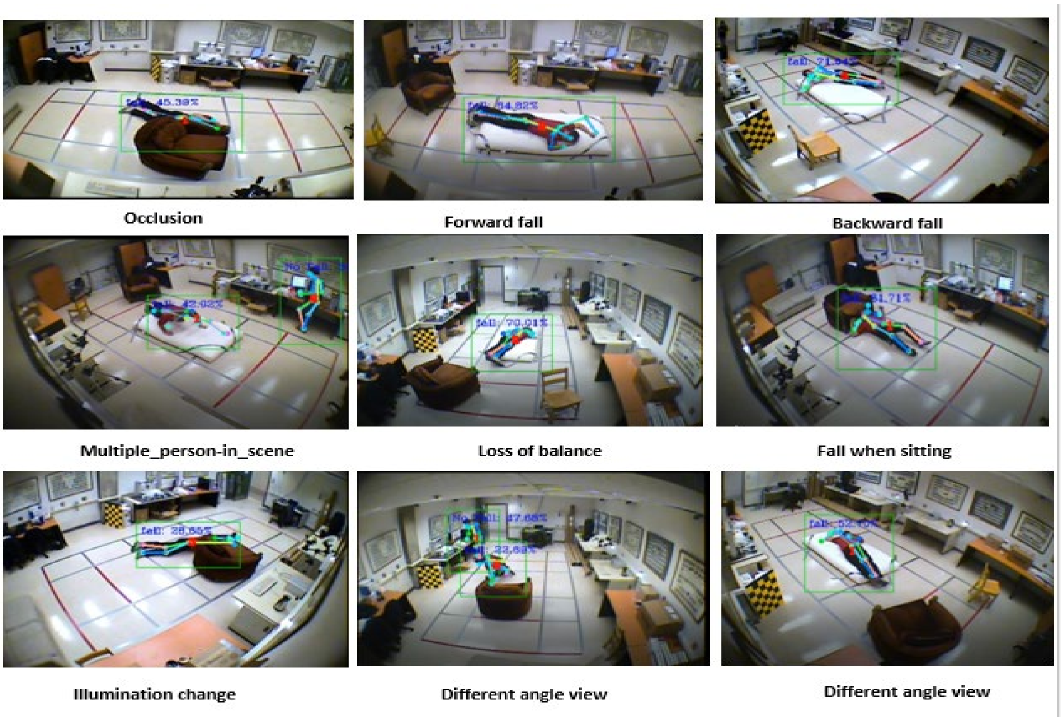
Multiple_person-in_scene

Different angle view

Different angle view

Illumination change

Fig. 10. The experimental results using the URFD dataset.


Occlusion

Forward fall

Backward fall

Multiple_person-in_scene

Loss of balance

Fall when sitting

Illumination change

Different angle view

Different angle view

Fig. ١١. The experimental results using the MCF dataset.

*5.*Conclusion

The proposed method presented aims to overcome the most significant issues with the visual fall detection system since it is crucial to reduce fall-related injuries and fatalities. The main idea for this method is to use the skeleton data for the human body and a convolutional graph network. The proposed method can recognize both temporal and spatial properties of detected human skeleton joints because it is based on graphical data that deals with skeletal information about people. The proposed method is based on a transfer learning technique that employs a pre-trained ST-GCN model. The first phase of the proposed method is to detect a human in the video frame using the Yolo algorithm, which works in real time with speed and accuracy. The second phase is extracting the skeleton data for the human body using the OpenPose technique, which can determine the position of the main body joints and represent them as (X, Y, C), where (X, Y) represent the node location and C represents the confidence score. The final phase is the classification task using the spatial-temporal convolution neural network (ST_GCN). The proposed method can detect a fall event accurately based on the human body information obtained from applying the open pose technique. The ST_GCN can obtain both temporal and spatial information for the extracted features. The three publicly available datasets used in this research, which are the most common and significant in this field, are used to test and evaluate the proposed method both quantitatively and qualitatively. The experimental results achieved high performance accuracy, with an average detection rate of 98.6% for the FDD dataset, 98.2% for the URFD dataset, and 97.5% for the MCF dataset. Under various situations (such as different illumination changes, camera angles, multiple people in the same scene, and different falling transitions), these results were obtained. When comparing the proposed method to previous fall detection techniques, it has high prediction accuracy. It also operates in real-time and protects user privacy, which makes it useful in enclosed spaces like hospitals and nursing homes. We may conclude from the above that the suggested method in this work is fully functional. Additionally, it is exposed and tested only in indoor places. Therefore, the system has to be optimized in order to function in places with more complicated settings. In the future, we'll try to apply our technique to multi-person tracking in the outdoors places and in other complicated datasets.

References

[1] Kubitza, Jenny, et al. "Therapy Options for Those Affected by a Long Lie After a Fall: A Scoping Review." BMC Geriatrics, vol. 22, no. 1, Springer Science and Business Media LLC, July 2022. Crossref, doi:10.1186/s12877-022-03258-2.

[2] Hallaj, Fatima. "Effect of Balance Exercise on Risk of Falls Among Institutionalized Elders in Lattakia, Syria." Alexandria Scientific Nursing Journal, vol. 20, no. 2, Egypts Presidential Specialized Council for Education and Scientific Research, Dec. 2018, pp. 81–96. Crossref, doi:10.21608/asalexu.2018.208195.

[3] Faulkner, Nathaniel, et al. "CapLoc: Capacitive Sensing Floor for Device-Free Localization and Fall Detection." IEEE Access, vol. 8, Institute of Electrical and Electronics Engineers (IEEE), 2020, pp. 187353–64. Crossref, doi:10.1109/access.2020.3029971.

[4] Bouwmans, Thierry, et al. "Decomposition into Low-rank Plus Additive Matrices for Background/Foreground Separation: A Review for a Comparative Evaluation with a Large-scale Dataset." Computer Science Review, vol. 23, Elsevier BV, Feb. 2017, pp. 1–71. Crossref, doi: 10.1016/j.cosrev.2016.11.001.

[5] D.P., Sangeetha. "Fall Detection for Elderly People Using Video-based Analysis." Journal of Advanced Research in Dynamical and Control Systems, vol. 12, no. SP7, Institute of Advanced Scientific Research, July 2020, pp. 232–39. Crossref, doi:10.5373/jardcs/v12sp7/20202102.

[6] Alam, Ekram, et al. "Vision-based Human Fall Detection Systems Using Deep Learning: A Review." Computers in Biology and Medicine, vol. 146, Elsevier BV, July 2022, p. 105626. Crossref, doi: 10.1016/j.compbiomed.2022.105626.

[7] Tripathi, Rajesh Kumar, et al. "Real-time Based Human-fall Detection from an Indoor Video Surveillance." International Journal of Applied Pattern Recognition, vol. 5, no. 1, Inderscience Publishers, 2018, p. 72. Crossref, doi:10.1504/ijapr.2018.10011655.

[8] Kim, Young-Min, et al. "CCTV Object Detection With Background Subtraction and Convolutional Neural Network." KIISE Transactions on Computing Practices, vol. 24, no. 3, Korean Institute of Information Scientists and Engineers, Mar. 2018, pp. 151–56. Crossref, doi:10.5626/ktcp.2018.24.3.151.

[9] Rahim, Robbi. "Internet of Things Based Driver Exhaustion Detection System Using Distributed Sensor Network." International Journal on Recent and Innovation Trends in Computing and Communication, vol. 8, no. 4, Auricle Technologies, Pvt., Ltd., Apr. 2020, pp. 12–16. Crossref, doi:10.17762/ijritcc. v8i4.5423.

[10] Mamchur, Nazar, et al. "Person Fall Detection System Based on Video Stream Analysis." Procedia Computer Science, vol. 198, Elsevier BV, 2022, pp. 676–81. Crossref, doi: 10.1016/j.procs.2021.12.305.

[11] Galvão, Yves M., et al. “A Multimodal Approach Using Deep Learning for Fall Detection.” Expert Systems With Applications, vol. 168, Elsevier BV, Apr. 2021, p. 114226. Crossref, doi: 10.1016/j.eswa.2020.114226.

[12] Rodriguez-Conde, Ivan, et al. “Optimized Convolutional Neural Network Architectures for Efficient On-device Vision-based Object Detection.” Neural Computing and Applications, vol. 34, no. 13, Springer Science and Business Media LLC, Dec. 2021, pp. 10469–501. Crossref, doi:10.1007/s00521-021-06830-w.

[13] GALOF, Katarina, and Nevenka GRIČAR. “Independent Living of the Elderly in the Home Environment.” International Journal of Health Sciences (IJHS), vol. 5, no. 2, American Research Institute for Policy Development, 2017. Crossref, doi:10.15640/ijhs.v5n2a2.

[14] Chen, Yong, et al. “Vision-Based Fall Event Detection in Complex Background Using Attention Guided Bi-Directional LSTM.” IEEE Access, vol. 8, Institute of Electrical and Electronics Engineers (IEEE), 2020, pp. 161337–48. Crossref, doi:10.1109/access.2020.3021795.

[15] Ge, Chenjie et al. “Human fall detection using segment-level cnn features and sparse dictionary learning.” 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP) (2017): 1-6.

[16] Núñez-Marcos, Adrián, et al. “Vision-Based Fall Detection with Convolutional Neural Networks.” Wireless Communications and Mobile Computing, vol. 2017, Hindawi Limited, 2017, pp. 1–16. Crossref, doi:10.1155/2017/9474806.

[17] Alanazi, Thamer, and Ghulam Muhammad. “Human Fall Detection Using 3D Multi-Stream Convolutional Neural Networks With Fusion.” Diagnostics, vol. 12, no. 12, MDPI AG, Dec. 2022, p. 3060. Crossref, doi:10.3390/diagnostics12123060.

[18] Tsai, Tsung-Han, and Chin-Wei Hsu. “Implementation of Fall Detection System Based on 3D Skeleton for Deep Learning Technique.” IEEE Access, vol. 7, Institute of Electrical and Electronics Engineers (IEEE), 2019, pp. 153049–59. Crossref, doi:10.1109/access.2019.2947518.

[19] Chen, Weiming, et al. “Fall Detection Based on Key Points of Human-Skeleton Using OpenPose.” Symmetry, vol. 12, no. 5, MDPI AG, May 2020, p. 744. Crossref, doi:10.3390/sym12050744.

[20] Wang, Rui, et al. “Global Relation Reasoning Graph Convolutional Networks for Human Pose Estimation.” IEEE Access, vol. 8, Institute of Electrical and Electronics Engineers (IEEE), 2020, pp. 38472–80. Crossref, doi:10.1109/access.2020.2973039.

[21] Zhao, Liquan, and Shuaiyang Li. “Object Detection Algorithm Based on Improved YOLOv3.” Electronics, vol. 9, no. 3, MDPI AG, Mar. 2020, p. 537. Crossref, doi:10.3390/electronics9030537.

[22] Cao, Zhe, et al. “OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields.” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, Institute of Electrical and Electronics Engineers (IEEE), Jan. 2021, pp. 172–86. Crossref, doi:10.1109/tpami.2019.2929257.

[23] Yan, Sijie, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition." Thirty-second AAAI conference on artificial intelligence. 2018.

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

[25] UR Fall Detection Dataset.” UR Fall Detection Dataset, fenix.ur.edu.pl/~mkepski/ds/uf.html.

[26] E. Auvinet, C. Rougier, J.Meunier, A. St-Arnaud, J. Rousseau, "Multiple cameras fall dataset", Technical report 1350, DIRO - Université de Montréal, July 2010."

[27] Keskes, Oussema, and Rita Noumeir. “Vision-Based Fall Detection Using ST-GCN.” IEEE Access, vol. 9, Institute of Electrical and Electronics Engineers (IEEE), 2021, pp. 28224–36. Crossref, doi:10.1109/access.2021.3058219.

[28] Yi, Dokkyun, et al. “An Adaptive Optimization Method Based on Learning Rate Schedule for Neural Networks.” Applied Sciences, vol. 11, no. 2, MDPI AG, Jan. 2021, p. 850. Crossref, doi:10.3390/app11020850.

[29] Ying, Xue. “An Overview of Overfitting and Its Solutions.” Journal of Physics: Conference Series, vol. 1168, IOP Publishing, Feb. 2019, p. 022022. Crossref, doi:10.1088/1742-6596/1168/2/022022.

[30] Andreieva, Valeria, and Nadiia Shvai. “Generalization of Cross-Entropy Loss Function for Image Classification.” Mohyla Mathematical Journal, vol. 3, National University of Kyiv - Mohyla Academy, Jan. 2021, pp. 3–10. Crossref, doi:10.18523/2617-7080320203-10.

[31] Zeng, Changchang, et al. “A Survey on Machine Reading Comprehension—Tasks, Evaluation Metrics and Benchmark Datasets.” Applied Sciences, vol. 10, no. 21, MDPI AG, Oct. 2020, p. 7640. Crossref, doi:10.3390/app10217640.

[32] Xu, Jianfeng, et al. “Three-way Confusion Matrix for Classification: A Measure Driven View.” Information Sciences, vol. 507, Elsevier BV, Jan. 2020, pp. 772–94. Crossref, doi: 10.1016/j.ins.2019.06.064.

[33] Alaoui, Abdessamad Youssfi, et al. “Fall Detection for Elderly People Using the Variation of Key Points of Human Skeleton.” IEEE Access, vol. 7, Institute of Electrical and Electronics Engineers (IEEE), 2019, pp. 154786–95. Crossref, doi:10.1109/access.2019.2946522.

[34] Harrou, Fouzi, et al. “An Integrated Vision-Based Approach for Efficient Human Fall Detection in a Home Environment.” IEEE Access, vol. 7, Institute of Electrical and Electronics Engineers (IEEE), 2019, pp. 114966–74. Crossref, doi:10.1109/access.2019.2936320.

[35] M. M. Hasan, M. S. Islam and S. Abdullah, "Robust Pose-Based Human Fall Detection Using Recurrent Neural Network," 2019 IEEE International Conference on Robotics, Automation, Artificial-intelligence and Internet-of-Things (RAAICON), 2019, pp. 48-51, doi: 10.1109/RAAICON48939.2019.23.

[36] Núñez-Marcos, Adrián, et al. “Egocentric Vision-based Action Recognition: A Survey.” Neurocomputing, vol. 472, Elsevier BV, Feb. 2022, pp. 175–97. Crossref, doi: 10.1016/j.neucom.2021.11.081.

[37] Chen, Yong, et al. “Vision-Based Fall Event Detection in Complex Background Using Attention Guided Bi-Directional LSTM.” IEEE Access, vol. 8, Institute of Electrical and Electronics Engineers (IEEE), 2020, pp. 161337–48. Crossref, doi:10.1109/access.2020.3021795.

[38] Wang, Bo-Hua, et al. “Fall Detection Based on Dual-Channel Feature Integration.” IEEE Access, vol. 8, Institute of Electrical and Electronics Engineers (IEEE), 2020, pp. 103443–53. Crossref, doi:10.1109/access.2020.2999503.

[39] Xu, Qingzhen, et al. “Fall Prediction Based on Key Points of Human Bones.” Physica A: Statistical Mechanics and Its Applications, vol. 540, Elsevier BV, Feb. 2020, p. 123205. Crossref, doi: 10.1016/j.physa.2019.123205.

[40] Carlier, Alexy et al. “Fall Detector Adapted to Nursing Home Needs through an Optical-Flow based CNN.” 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology

[41] Mamchur, Nazar, et al. “Person Fall Detection System Based on Video Stream Analysis.” Procedia Computer Science, vol. 198, Elsevier BV, 2022, pp. 676–81. Crossref, doi: 10.1016/j.procs.2021.12.305.