Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.042869

ARTICLE

Design of a Lightweight Compressed Video Stream-Based Patient Activity Monitoring System

Sangeeta Yadav[1], Preeti Gulia[1,*], Nasib Singh Gill[1,*], Piyush Kumar Shukla[2], Arfat Ahmad Khan[3], Sultan Alharby[4], Ahmed Alhussen[4] and Mohd Anul Haq[5]

[1]Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, 124001, India

[2]University Institute of Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, 462033, India

[3]Department of Computer Science, College of Computing, Khon Kaen University, Khon Kaen, 40002, Thailand

[4]Department of Computer Science, College of Computer and Information Sciences, Majmaah University, Al Majmaah, 11952, Kingdom of Saudi Arabia

[5]Department of Computer Engineering, College of Computer and Information Sciences, Majmaah University, Al Majmaah, 11952, Kingdom of Saudi Arabia

*Corresponding Authors: Preeti Gulia. Email: preeti@mdurohtak.ac.in; Nasib Singh Gill. Email: nasib.gill@mdurohtak.ac.in

Received: 15 June 2023 Accepted: 25 September 2023 Published: 30 January 2024

ABSTRACT

Inpatient falls from beds in hospitals are a common problem. Such falls may result in severe injuries. This problem can be addressed by continuous monitoring of patients using cameras. Recent advancements in deep learning-based video analytics have made this task of fall detection more effective and efficient. Along with fall detection, monitoring of different activities of the patients is also of significant concern to assess the improvement in their health. High computation-intensive models are required to monitor every action of the patient precisely. This requirement limits the applicability of such networks. Hence, to keep the model lightweight, the already designed fall detection networks can be extended to monitor the general activities of the patients along with the fall detection. Motivated by the same notion, we propose a novel, lightweight, and efficient patient activity monitoring system that broadly classifies the patients' activities into fall, activity, and rest classes based on their poses. The whole network comprises three sub-networks, namely a Convolutional Neural Networks (CNN) based video compression network, a Lightweight Pose Network (LPN) and a Residual Network (ResNet) Mixer block-based activity recognition network. The compression network compresses the video streams using deep learning networks for efficient storage and retrieval; after that, LPN estimates human poses. Finally, the activity recognition network classifies the patients' activities based on their poses. The proposed system shows an overall accuracy of approx. 99.7% over a standard dataset with 99.63% fall detection accuracy and efficiently monitors different events, which may help monitor the falls and improve the inpatients' health.

KEYWORDS

Fall detection; activity recognition; human pose estimation; accuracy

1 Introduction

Inpatient falls in hospitals from the beds are a serious problem and sometimes lead to severe injuries like bone fractures and soreness, eventually resulting in worsening patients' health conditions. These events are commonly observed in the psycho-geriatric wards and mental health hospitals, where patients suffering from vertigo, dizziness, and cognitive impairment are admitted. Such events may bother the medical staff associated with them and result in guilt, anxiety, and sometimes litigation. However, the hospital staffs have their protocols to avoid or prevent these falls. The advancements in intelligent systems in generating timely alerts may aid in serving better and improved healthcare.

Several modern hospitals use deep learning and computer vision-based non-invasive in-clinic patient monitoring systems to monitor and observe patient behaviour. Such vision-based monitoring systems have performed well in monitoring different poses and behaviour of the patients, like breathing disorders [1], epileptic patients [2], sleeping postures [3], walking activity [4], and infant motions. The recent advancements in fall detection using camera-captured video streams and fall prediction systems [5] (detecting a fall when it happened rather than predicting before when it was about to happen) have significantly improved their performance over standard synthetic libraries or simulated datasets.

However, the estimation of human poses for the inpatient analysis has gained much attention, but its usage for patient fall prevention from beds is less explored. Many systems have been designed to detect the patient's lying position relative to the edge of the bed. Some systems focused on seeing the patient's behaviour during bed exits. For predicting falls in such scenarios, the in-clinic monitoring systems continuously monitor the patient's essential motion or pose. Some systems employ pressure pads to detect patients on/off bed position [6], but their applicability shows several false alarms, resulting in alarm fatigue. The vision-based monitoring systems detect the action regarding exit from the bed or sitting position from the images or video stream captured by the cameras. Though several patients' falls also happen when they try to get out of bed, such falls are also associated with additional risks like restless sleep, uncontrolled motions, or unusual dreams that may trigger them to jump out of bed for protection. Such activities also add to the more patient falls and are considered in this paper.

The non-invasive fall detection or gait recognition systems may be broadly classified into two categories, i.e., vision or camera-based systems or wearable sensor-based systems. Pressure sensors, gyroscopes, accelerometers, microphones, etc., are used as wearable sensors to detect the change in acceleration or location of the patient. Then, these parameters are used for fall detection. Recent developments led to performance improvement in fall detection or gait recognition systems using wearable sensors [7–9]. Several sensor-based applications have also been developed. But the patients are unwilling or feel uncomfortable wearing those sensors all day.

Moreover, the factors like noise or some routine activities like sitting or lying hampered the performance of these devices and resulted in false alarms. These limitations attracted the attention of researchers toward vision-based monitoring systems as they are free from inconvenience. The recent advancements in computer vision using deep learning approaches have led to the exploration of many vision-based patient monitoring systems. The accuracy of such systems has improved over the years, but some challenges, like complex backgrounds, lighting variation, etc., which may lead to false alarms, also need to be addressed or explored more. Moreover, deep learning-based vision systems have high computation costs, which limit their real-time performance. Achieving high accuracy with lower computation costs is the avenue of research in this domain.

In the computer vision domain, one of the fundamental tasks is human pose estimation, and it deals with localizing the critical points of human position in a 3D space or an image. Human pose estimation substantially contributes to applications like animation, computer interaction, action

recognition, etc. Recent advancements in 2D pose estimation [10–12] have significantly improved the accuracy of these systems over massive datasets. Moreover, some notable performance improvements were also observed in 3D human pose estimation systems [13,14]. The complex background or lighting variations also significantly impact the accuracy of fall detection tasks. But using human poses has substantially improved these systems' generalization ability and accuracy by addressing these challenges to some extent.

Although the recent advancements in vision-based systems have achieved detection accuracy up to a remarkable level but with the enhanced computation cost [15], the extracted features are not robust enough to deal with complex or challenging conditions. Also, the fall period varies for different patients or age groups [16]. Generally, the fall of aged persons lasts longer than younger ones, and activities like lying on the bed differ from falls. Such challenges are needed to be addressed. The input must comprise more video frames, as some works [17] based on short video streams resulted in false detections.

The developments in vision based-detection tasks during the last few years have been discussed above. Vision-based systems with deep learning approaches have high computation costs. Hence wearable systems are designed to monitor and observe the patient's behaviour, for example, fall detection, breathing disorders, etc. Wearable-based monitoring approaches require the patient to wear sensors or other monitoring equipment on their body or clothing, which can be invasive [9,18,19]. Sometimes, it can be uncomfortable for the patient. There is no general-purpose non-invasive monitoring system that identifies the general activities of the patients to monitor their well-being and improvements without their direct contact. Hence, a lightweight video-based design is required, which will detect or monitor the general activities of the patients along with fall detection. Motivated by the same notion, a lightweight patient activity monitoring system has been proposed, which broadly monitors the patients' events in three categories: falls, rest, and activity. It is the expansion of the existing detection networks for monitoring tasks without incurring additional high computational costs by sharing the same network with rest and activity detection tasks. For the classification task, Classical K-Nearest Neighbors (KNN), Random Forest (RF), and Decision Tree algorithms, in addition to a You Only Look Once (YOLO), the researchers offered neural networks and several others. Scientists used machine learning techniques, CNN, and deep algorithm development to create methods for recognizing, classifying, and differentiating various poses [20]. The proposed network may also provide accounting information by installing corresponding counters.

The main focus of this work is to improve the solution's value by expanding the activity detection without impacting the accuracy of fall detection and increasing computational costs. The whole process is carried out in three significant steps. Firstly, a video compression network efficiently stores the raw video frames in compressed format. Its sub-networks frame autoencoder, flow autoencoder, and motion extension networks comprise deep learning layers. Secondly, 3D poses are estimated from the video. Then, falls are recognized from these poses. 3D pose estimation is formulated as 2D and then lifted to 3D. And to keep the system lightweight and less computative, a 2D pose estimator and a lifting network are employed. Thirdly, a 3D pose of each frame is given as input to the activity recognition network.

The main contributions of this work are listed below:

- Firstly, the proposed system consumes the encoded video streams to classify a patient's poses into fall, activity and rest classes.
- Secondly, it is a lightweight patient activity monitoring system because the already designed fall detection network has been extended to monitor the activity, rest events, and fall events. It will

help in monitoring the general well-being of the patients. Moreover, rest and activity time can also be accounted for by using corresponding counters.

- Thirdly, the system is an amalgamation of three networks: a CNN-based video compression network, an LPN for human pose estimation, and a ResNet Mixer block-based activity recognition network. The experiment uses a standard Nanyang Technological University's Red Green Blue (NTU RGB) dataset. The performance values show that the system delivers an overall accuracy of 99.77% with 99.63% fall detection accuracy.

The rest of the paper is organized as follows: Section 2 elaborates on the related works in fall detection and patient monitoring systems. In Section 3, the proposed system's details are described step-wise. The details of the dataset and the evaluation parameters used in the implementation are also defined. The experiment design and performance results of the proposed system are covered in Section 4. A comparative analysis of the proposed method and existing approaches has also been presented. The conclusion of this research work is given in Section 5.

2 Literature Review

The various techniques of vision-based fall detection, fall risk assessment systems, and human pose estimation have been thoroughly reviewed and analyzed. Several fall detection systems based on sensors have been proposed and experimented. They have employed diverse classification algorithms [21,22]. Some techniques also differ in the sensors used, like infrared cameras, Kinect, depth cameras, and RGB cameras. One prominent approach to fall detection using Red, Geen and Blue (RGB) frames has been proposed by Lu et al. [23]. It comprises Long Short Term Memory (LSTM) and 3D CNN to deal with insufficient fall information. In another approach, a depth camera is used to attain the 3D coordinates of the patient's key points, and then the fall is detected by classifying these poses [24]. Thermal sensors are also used in the multi-occupancy space to provide an end-to-end solution within the environment [25]. Nowadays, the easier retrieval of 3D data led to the wide use of Kinect in these fall prediction systems. But as the Kinect's depth camera has limited coverage, it is suitable for only small spaces.

The different fall actions can be classified using Support Vector Machine (SVM) or decision trees. But the, Deep Neural Networks (DNNs) have more accuracy than these techniques with fewer feature engineering chores. One such CNN-based fall detection approach has been proposed, recognizing fall events and daily life activities using the CNN network. Furthermore, 3D CNN-based networks are designed to improve the accuracy of these systems by exploiting temporal and spatial information. 3D CNN, in combination with Variational Auto-encoder and region extraction approach, is also used for performance enhancement in fall detections. One promising traditional algorithm-based technique has also been proposed by Tsai et al. [24]. This approach transforms the depth information into 3D poses, and the fall events are detected using 1D CNN. The depth cameras are employed in this technique, and the 3D poses are recognized directly by the human pose estimator from the RGB frames. This approach makes the system free from limited distance coverage. Several experiments have also been conducted to assess the performance of different neural networks over fall detection tasks [26,27].

The human pose estimator localizes the human key points. Based on the output requirements, they can be categorized into 3D or 2D pose estimation. Deep learning evolved as a potential tool for 2D pose estimation. Several machine learning-based techniques have been developed to identify and classify human gait [19]. Joint heatmaps are used to train the signals for performance improvement. A U-shaped network is also proposed to refine the prediction by integrating diverse hourglass modules. Cao et al. [11] proposed a system that works for real-time multi-person pose estimation. Firstly, this

system identifies the human key points and then integrates them into a person using corresponding part affinity fields. HRNet, a multi-scale fusion network with high-resolution feature maps and a top-down pipeline, performs better.

The 3D human poses estimation deals with localizing human vital points from the videos or images in 3D space. The earlier 3D estimation methods used deep neural networks to predict 3D coordinates directly. Though the feature sets about a particular task are well learned in these systems but with enhanced complexity and high computation cost. After introducing 2D pose estimation, the approaches based on 2D estimates become more prevalent. The complexity of these models is reduced to an extent by direct depth prediction of each key point. The depth information is predicted by taking the concatenation of 2D poses as input. Weakly supervised learning is used to train the 3D models by employing the re-projection method to deal with insufficient data in 3D estimations. The human poses are modified using video information to avoid false pose estimations [28], and the occlusion problem is avoided using multi-view images. One of the efficient pose-based detection works using the videos has been proposed by Chen et al. [29], but this work is limited to fall detection tasks only. The outcomes discussed above are also summarised in below Table 1.

Table 1: Summary of some of the prominent works on fall detection and activity recognition

Paper	Network/Techniques used	Specifications
[11]	U-shaped network	Refine the prediction by integrating diverse hourglass modules
[19]	Machine Learning Methods (K-Nearest Neighbour (KNN), SVM, Extreme Learning Machines (ELM), Multi-Layer Perceptron (MLP))	For identification and classification of human gait
[21]	LSTM Neural Networks	Fall detection only
[22]	AlphaPose, Random Forest (RF), SVN, MLP, KNN	Vision-based Fall detection and Activity Recognition using Human Pose Estimation
[23]	LSTM, along with 3D CNN	Fall detection using RGB frames
[24]	Highly robust deep convolution neural network	Uses a pruning method to reduce parameters and calculations of the fall detection network
[25]	The vision-based approach of multi-occupancy fall detection (MoT-LoGNN)	Thermal vision-based discrimination between a fall or no-fall
[26]	CNN	Fall detection only
[27]	Multi-Scale Skip Connection Segmentation Network (MSSkip), LSTM	Detect fall or no fall
[28]	Multi-stride Temporal Convolutional Networks (TCNs)	Assess whether the predicted pose forms a valid posture and a valid movement

(Continued)

Table 1 (continued)

Paper	Network/Techniques used	Specifications
[29]	Pyroelectric Infrared Sensors (PIR), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), Naïve Bayes (NB), and AdaBoost (AB)	Detection of regular motions and patient fall events

The various Kinect-camera-based datasets widely used in fall detection and activity recognition works are summarised in Table 2.

Table 2: The datasets widely used for fall detection tasks

Reference	Dataset	Fall/Activity	ML technique	Accuracy/F1-score
[30]	SDU Fall	Fall to the floor, Sitting, walking, squatting, lying, bending	Bag of words model built upon curvature scale space feature	Accuracy: 79.91%
[31]	EDF-OCCU	Falls in 8 different directions, Laying, picking up, sitting (floor), tying shoelaces, plank exercise		
[32]	UR Fall	From standing, sitting on a chair, Lying, walking, sitting down, crouching down	Support vector machine, k-nearest neighbour	Accuracy: 94.99%,
[33]	TST v2	From the front, backwards, to the side, Sitting, grasping an object from the floor, walking, lying	Depth frame	
[34]	PKU-MMD	Drinking, waving hands, putting on the glassed, hugging, shaking	Regression RNN Regression SVM Bidirectional LSTM S-T attention LSTM	F1-score: 52.6% 13.1% 33.3% 31.6%
[35]	NTU RGB-D	Fall to the floor and 120 actions	S-T LSTM Part-Aware LSTM RNN CNN FSNet	Accuracy: 57.9% 26.3% 44.9% 61.8% 62.4%

Classifier models are a type of machine learning model used to classify or categorize data into different classes or categories. These models learn from labelled training data and use that knowledge to make predictions or assign labels to unseen or unlabeled instances. In addition to medical imaging, classifier models are widely used in domains such as image recognition, natural language processing, fraud detection, sports, and many others [36–38]. Some widely used classifier models are as follows:

- Logistic Regression: Logistic regression is a popular linear classifier that models the relationship between the input features and the probability of belonging to a particular class. It works well when the decision boundary is linear or can be approximated by a linear function.

- Naïve Bayes: Naïve Bayes classifiers are based on Bayes' theorem and assume that features are conditionally independent given the class. Despite the naive assumption, they often perform well and are particularly effective in text classification and spam filtering tasks.
- Decision Trees: Decision trees are hierarchical structures that recursively partition the data based on feature values. Each internal node represents a test on a particular feature, and each leaf node represents a class label. Decision trees are intuitive, easy to interpret, and can handle categorical and numerical data.
- Random Forest: Random forest is an ensemble learning method that combines multiple decision trees. It constructs a set of decision trees using random subsets of the training data and features, and the final prediction is made by aggregating the predictions of individual trees. Random Forests are robust, handle high-dimensional data well, and can capture complex interactions between features.
- K-Nearest Neighbors (KNN): KNN is a non-parametric classifier that classifies instances based on their proximity to labelled examples. It assigns the class label of most of the k nearest neighbours in the training data. KNN is simple to implement but can be computationally expensive for large datasets.
- Neural Networks: Neural networks, profound learning models, have gained significant attention in recent years. They consist of multiple interconnected layers of nodes (neurons) that learn hierarchical representations of the data. Neural networks can handle complex patterns and large amounts of data and are particularly effective for image and speech recognition tasks.

Some potential research gaps observed in the above-discussed research works are listed as under:

- Unobtrusiveness and user acceptance: Wearable fall detection devices should be discreet and comfortable to encourage long-term use and user acceptance. More research is needed to develop non-intrusive solutions that users are willing to adopt willingly.
- False alarm reduction: One of the challenges in fall detection systems is minimizing false alarms, as they can lead to user dissatisfaction and reduced trust in the system. Further investigation into advanced algorithms and sensor fusion techniques may help improve the accuracy of detecting actual falls while reducing false positives.
- Privacy and data security: As fall detection systems often collect sensitive health data, ensuring robust privacy and data security measures is crucial. Research should focus on developing methods to protect user data while maintaining system performance.
- Cost-effectiveness and scalability: Commercially available fall detection systems can be expensive, limiting their accessibility to a broader population. Future research should focus on developing cost-effective solutions that can be easily scaled for widespread adoption.
- Integration with healthcare systems: Investigate how fall detection systems can seamlessly integrate with healthcare providers' systems to provide timely alerts and support, leading to more efficient and effective healthcare interventions.
- Long-term monitoring and outcomes: Assess the impact of fall detection systems on overall fall rates, fall-related injuries, and healthcare outcomes over extended periods to understand their effectiveness and benefits in the long run.

Considering the various challenges and limitations of different fall detection networks, as discussed above, a lightweight and efficient video-based patient activity monitoring system for detecting inpatient fall and other activities has been proposed in this paper. The activity recognition network of the system sees inpatient falls using pose estimation and then timely alert the staff if a fall is detected.

This network has also been extended to monitor the patient's activity and rest time based on their different body positions.

3 Proposed Patient Monitoring System

The proposed patient monitoring system for inpatient fall reduction comprises a video compression network, a human pose estimator, and an activity recognition network. The video streams are compressed using the compression network and then stored in the compression format. The system fetches and regenerates the video frames and feeds them as input to the human pose estimator. It identifies the respective 3D human poses and provides them to the activity recognition network to further classify the actions into fall detection and activity and rest time estimation. If the network detects any imminent patient fall, instant alerts are made. The high-level structure of the proposed system is given in Fig. 1.

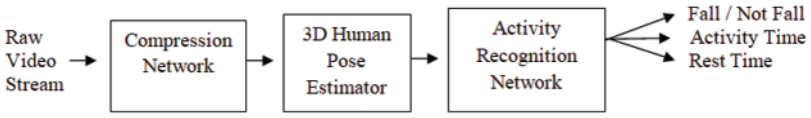

Figure 1: Various steps in the patients' activity monitoring system

3.1 Video Compression

The input video stream is compressed using an already-designed end-to-end video compression network [39]. The video compression network of the model is designed to compress the video frames so that classification can be made efficiently from the compressed format. It primarily comprises three sub-networks: Flow Autoencoder, Motion Extension Network, and Frame Autoencoder. The Frame Autoencoder is employed to compress the video frames. The input video frames are given as input to the encoder of the frame autoencoder, and it encodes them into binary format. The encoder comprises one Convolutional Gated Recurrent Unit (ConvGRU) layer and four 2D-CNN layers. ConvGRU layer is incorporated to enhance the compression quality as it encompasses the properties of both CNN and Recurrent Neural Networks (RNN). The binary format is quantized and then fed into the decoder for reconstruction. The difference between the original and the reconstructed frames gives the residual frame. The residual frame is then provided iteratively to the autoencoder to obtain the compressed format of the required compression quality. The network exploits multiple emissions to produce frames of varying compression quality. The structure of the compression network is given in Fig. 2.

The frame autoencoder employing recurrent ConvGRU-based encoder and decoder networks has been used to reconstruct the frames with varying degrees of compression quality. A ConvGRU-based encoder network encodes the frames in binary format, and then, after quantization, the corresponding decoder network regenerates the video frames according to the varying degree of compression. The motion information among the consecutive frames is computed efficiently using the Farneback-based flow estimation method. The calculated flow vector is fed into the flow encoder network for compression. The Frame Reconstruction Network is designed to reconstruct the current video frame. This network uses three inputs, i.e., regenerated structure, regenerated flow vector, and the previous frame, to estimate the current video frame. This network comprises convolutional layers only, and Rectified Linear Units are used as the activation functions. The detailed structure of this network is given in Fig. 3.

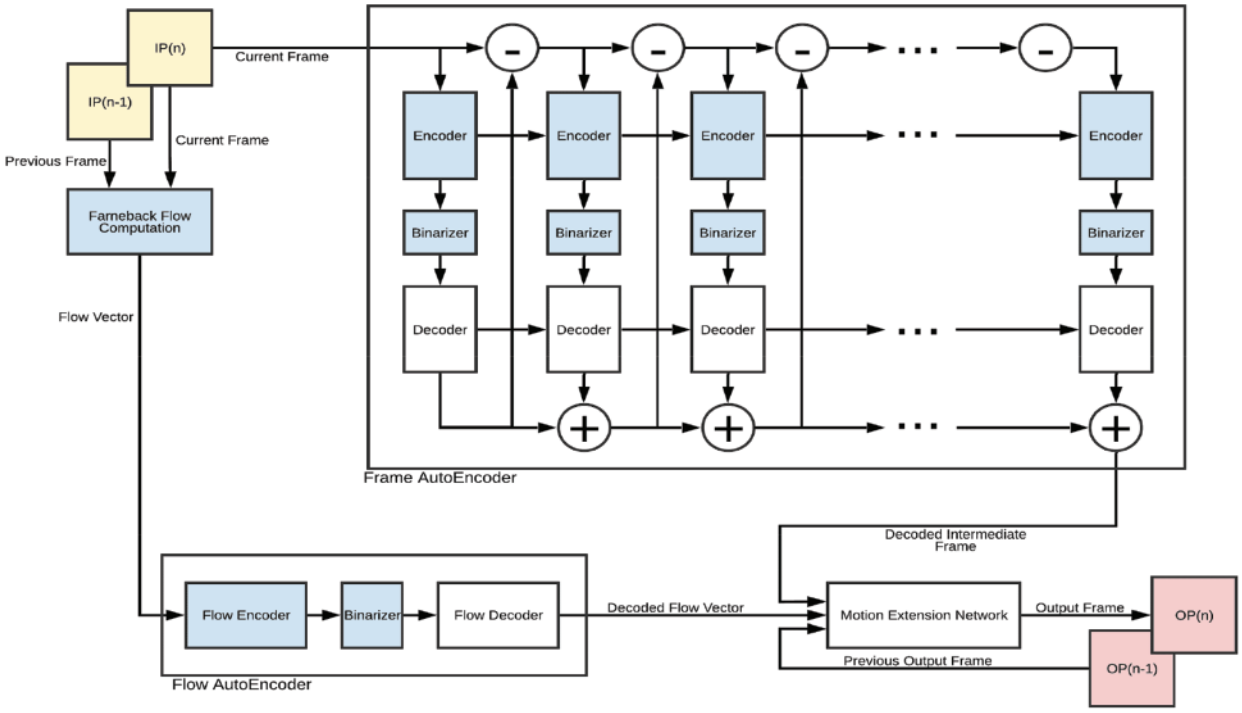

Figure 2: The video compression network

The intermediate representation is merged on the decoded middle frame, producing a high-quality current structure. The same has been represented by Eq. (1).

$$I_{decoded} = f_{decoder}\left(I_{encoded},\ F_{encoded},\ I_{decoded_{prev}}\right) \tag{1}$$

where "$I_{encoded}$ *and* $F_{encoded}$ *are binary encoding of current frame and flow vectors,* $I_{decoded_{prev}}$ *is previously decoded frame and* $f_{decoder}$ *is a representation of decoder neural network.*"

The video frames are then fed to the human pose estimator for further computations.

3.2 3D Human Pose Estimation

The proposed activity recognition network has been trained in a step-wise manner. For the 3D human pose estimation, firstly, 2D poses are predicted, which are then lifted to 3D. An example of a 2D pose and its corresponding 3D pose is given in Fig. 4.

As Lightweight Pose Network (LPN) is well-tuned with the taken NTU RGB + D dataset [30] and already trained on the Common Objects in Context (COCO) dataset [31], we have used LPN for 2D pose estimation from the video stills. Moreover, LPN achieves good accuracy with lesser complexity. According to the annotations, the joint heatmaps are produced, and the corresponding 2D coordinates are obtained by directly calculating the centre of mass of those heatmaps. These 2D poses are then normalized and scaled before lifting to 3D.

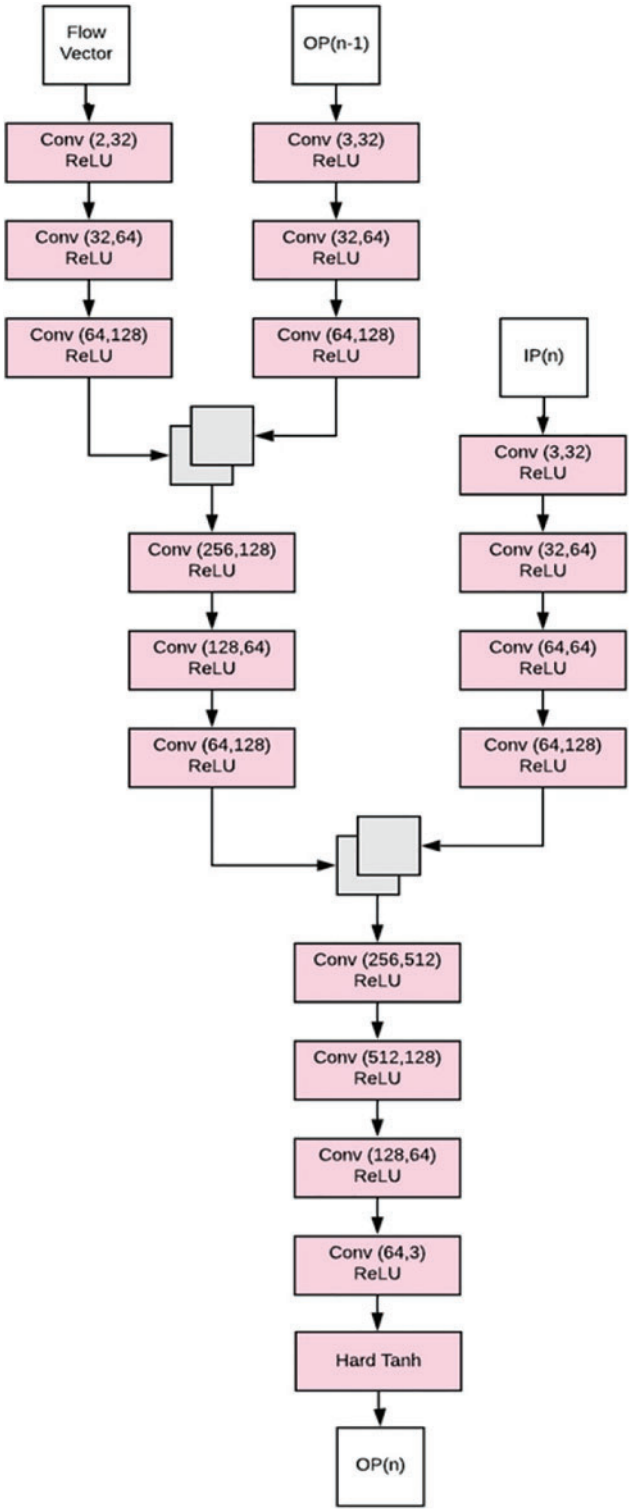

Figure 3: Video frame reconstruction network

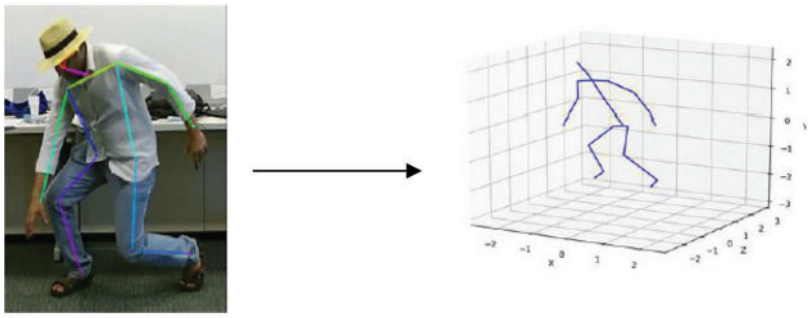

Figure 4: An example of the 2D and 3D pose

The 2D and 3D human poses are represented by x and y, respectively. 2D human poses, $x \in R^{2J,}$ are given as input to the lift network, which computes the corresponding 3D poses, $y \in R^{3J}$. The following functions, (2) and (3), minimize the prediction error.

$$f^* : R^{2J \rightarrow} R^{3J} \tag{2}$$

$$f^* = \min_f \frac{1}{N} \sum_{i=1}^{N} L\left(f\left(x_i\right) - y_i\right) \tag{3}$$

The mapping function f uses a 2 J vector for 2D pose x as input and a 3 J vector for 3D pose y as output. A depth dimension of J points is added to the 2D pose. 'i' represents the i^{th} pose among N poses.

The prediction error between the ground truth and the predicted pose is computed by using the following function:

$$L\left(\hat{y}_i, y_i\right) = \left|\left|\hat{y}_i - y_i\right|\right|^2 \tag{4}$$

where $\hat{y}_i$ and y_i represents the ground truth position and prediction of the i^{th} pose.

The structure of the lift network is given in Fig. 5. Along with convolutional layers, Batch Normalization (BN) layers are used for regularisation, and Dropout layers with a 0.5 value are used for more robustness.

3.3 Activity Recognition Network

The activity recognition network primarily detects or recognizes three activities: patient fall detection, rest, and action. The fall detection class recognizes the occurred fall and buzzers the associated alarm if a fall is detected. Similarly, the activity and rest poses are also identified based on their 3D poses. The whole process has been presented in Figs. 6 and 7. The counters associated with these classes may be used to monitor their corresponding metrics, i.e., for how long a patient has taken rest or performed any activity.

Earlier works utilize either LSTM or CNN-based modelling approaches for activity recognition networks. LSTM networks typically get current frame input at a time and need to remember it for long-term association. It makes long-term learning and feature association difficult. CNN networks address this issue by performing convolution on multiple frames windows. With a typical kernel, size ranges between 3 to 5, and one layer of CNN can form an association in a 3–5 frame window. This receptive field of the layer increases as the depth increases. It allows the final assimilation of all frames

in deeper layers. The receptive field can be improved faster by increasing the convolution kernel size in the time step direction. However, this significantly increases learning parameters and makes the model prone to overfitting.

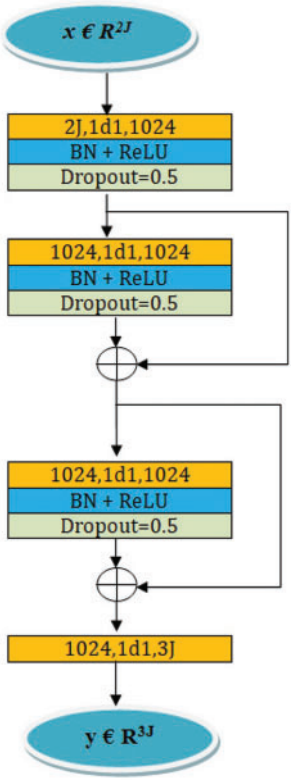

Figure 5: The structure of the lift network

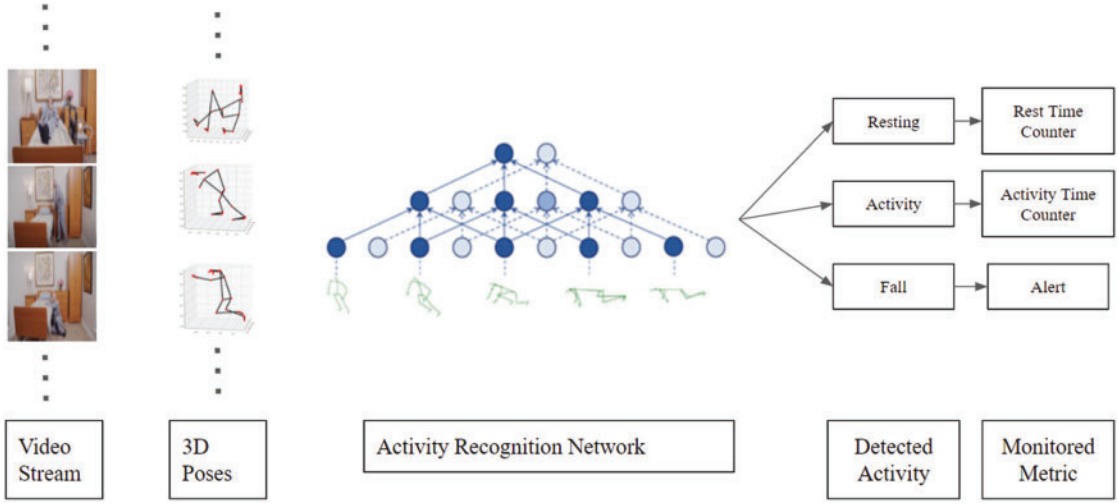

Figure 6: Recognition of patient fall and different activities by the proposed system

Figure 7: The joint aggregations corresponding to different activities

We propose an activity recognition network with a novel ResNet Mixer block to address the above issue. ResNet Mixer block comprises two 1D convolution layers, a BatchNorm layer, and a residual connection. It takes feature input of size – batch $*$ pose dimension $*$ the number of frames. We have kept the dimension and frame dimension to 148 to simplify the network. ResNet Mixer block takes out both temporal and spatial features from the input data to more effectively capture the relationships and patterns present in the data. The Mixer block operates by first processing the input data in the pose dimension, which involves applying the first 1D convolution layer to extract features from all joint position features of each frame. The Mixer block then processes the input data in the time dimension by applying a second 1D convolution to compute a new feature value for each feature based on its association in all 148 frames. BatchNorm and residual connection adds gradient stability to the network, allowing better training of deeper layers. It will enable the Resnet Mixer block to have a receptive field of all 148 frames. Multiple blocks allow iterative complex feature computation for activity detection with few parameters.

The Conv1D-based model employs a novel ResNet Mixer block as its core building block. The ResNet Mixer block comprises two 1D convolution layers, a BatchNorm layer, and a residual connection. The model processes the input data in pose and time dimensions to capture spatial and temporal features effectively. It aims to improve feature association and long-term learning, overcoming some limitations of traditional LSTM-based models. The ResNet Mixer block operates in two steps. Firstly, it processes the input data in the pose dimension, applying the first 1D convolution layer to extract features from all joint position features of each frame. It helps in capturing the spatial characteristics of the poses. Next, the ResNet Mixer block processes the input data in the time dimension, applying a second 1D convolution layer to compute a new feature value for each feature based on its association in all frames. This step captures temporal patterns and relationships among different structures. The architecture uses Multiple ResNet Mixer Blocks with Batch Normalization and Residual Connection and is designed to have a receptive field of all 148 frames.

The activity recognition network expands the number of pose features from 48 to 148 in the first convolution. It reduces the number of frames from 296 to 148 to make it compatible with the ResNet Mixer block. Following this, ten ResNet Mixer blocks are used. The advantage of using this network is that it has more depth and is less prone to overfitting. This network has 4,80,411 parameters, comparatively less than earlier developed networks (4.2 M parameters in the CNN model proposed by Cheng et al. [29]). The other details of the activity recognition network are as follows: 0.05 MB of Input size, 10.53 MB of Forward and backward pass size, 1.83 MB of Params size, and the anticipated Total size of 12.42 MB. The design of the individual component and the activity recognition network is given in Figs. 8a and 8b, respectively.

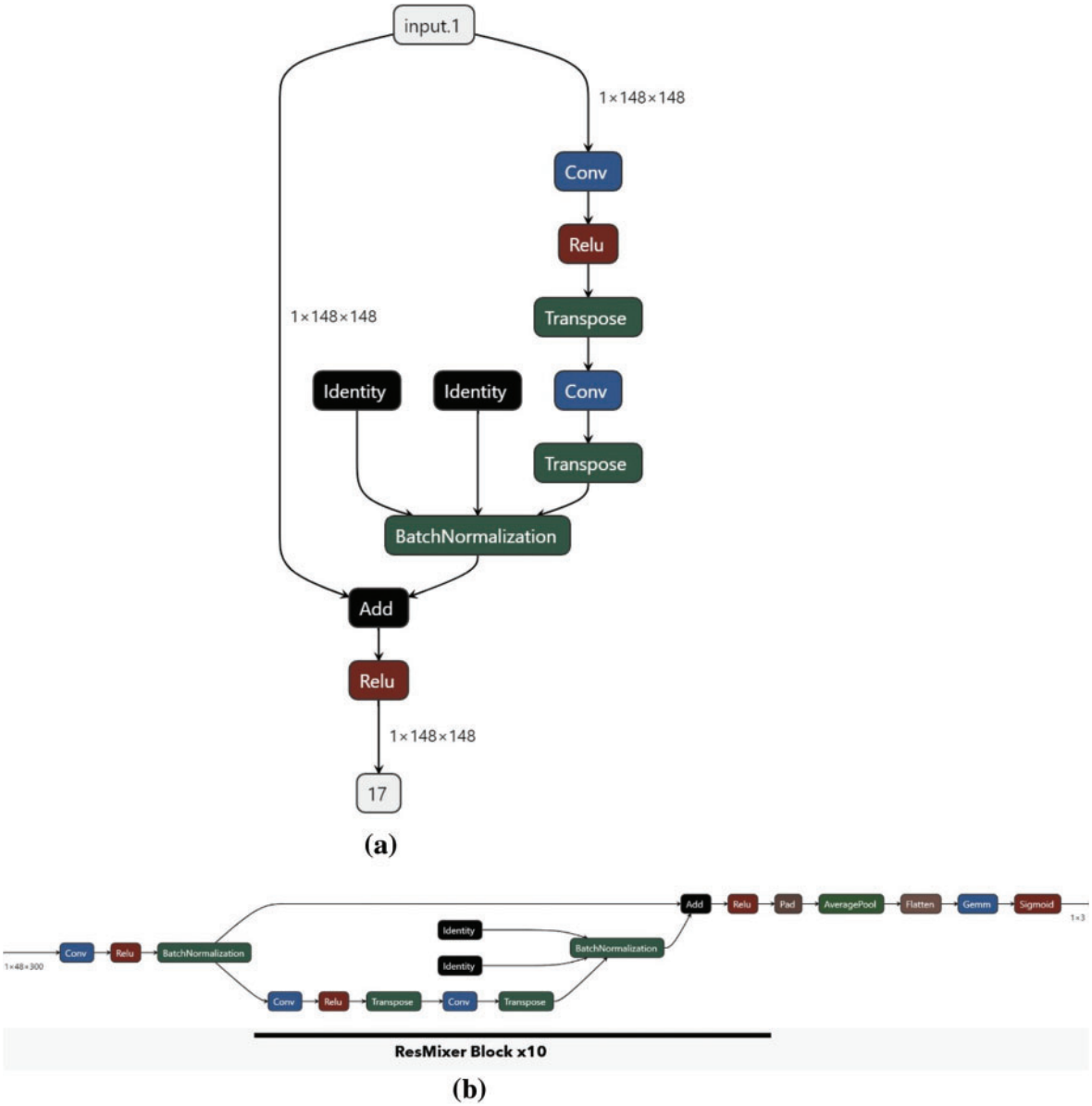

Figure 8: (a) Design of individual ResNet Mixer block. (b) The activity recognition network

4 Experiments and Analysis

In this section, the details of the experimentation have been covered. The datasets used and the training strategy of the designed network has been explained in detail.

4.1 Dataset

The video-based standard dataset NTU RGB+D has been used to train the network and assess the performance of the proposed system. This dataset comprises a total of 56,880 videos having 60 different action classes. The video sequences containing more than one person or missing poses or frames have been ignored, so only 44,372 video samples are used. Among them, 1490 video samples include fall events and resting events. Three cameras are used to obtain the videos of the dataset with angles, i.e., −45◦, 0◦, and +45◦. This data set also contains their infrared videos and 3D skeletal data, and each sample's depth map sequences and RGB videos. The 3D skeletal data comprises 3D coordinates of twenty-five joints. 1920×1080 is the resolution of RGB frames with a speed of 30 per second (FPS).

The video sequences containing more than one person or missing poses or frames have been ignored in the data preprocessing step. The dataset provides 890 videos for the fall category. We used 600 videos from the reading and writing category for the resting event with the person sitting on the chair. The division of skeleton data with its Frobenius norm is used for scaling purposes, and centring to its root joint gives its normalization. We used the standard weighted loss function to address the imbalance issue in the dataset. We use a large batch size (320) to ensure that positive samples are stochastically present in every batch.

Loss Function: The goal of the network is to reduce the structural distortion between the input video and the output. We also used the mean squared error (MSE) loss function to minimize colour distortion in decompressed images.

$$L = L_{ssim} + \alpha\, L_{mse} \tag{5}$$

where MSE error is evaluated as:

$$L_{mse}(y, y') = \frac{1}{N}\sum_{0}^{n}\left(y - y_i'\right)^2 \tag{6}$$

and SSIM error is evaluated based upon three comparison measurements, luminance (l), contrast (c) and structure (s):

$$L_{ssim}(y, y') = [l(y, y') \cdot\ c(y, y') \cdot\ s(y, y')] \tag{7}$$

4.2 Implementation Details

For the implementation purpose, a single T4, K80, or P100 GPU has been used to train the network on the Google Colaboratory platform. The frames have been kept to the size of 64×64. Adam Optimizer prepares the neural network, with 10e-4 being the learning rate. ג1 is taken as one, and ג2 be 10. The first frame encoder is trained for 100 epochs. Then the complete network is trained end-to-end for 70 epochs. During the training of the frame encoder with 100 epochs, the learning rate was divided by ten at the 50th, 70th, and 90th epoch. For the whole model training, only 70 epochs have been used after stacking the pre-trained framer encoder first, and the learning rate has been altered by dividing by ten at the 35th and 55th epochs.

4.3 Training Details

The proposed activity recognition network has been trained in a step-wise manner. For the 3D human pose estimation, firstly, 2D poses are predicted, which are then lifted to 3D. LPN is well-tuned with the taken NTU RGB + D dataset [40] and already trained on the COCO dataset [41], so we have used LPN for 2D pose estimation from the video stills. According to the annotations, the joint heatmaps are produced, and the corresponding 2D coordinates are obtained by directly calculating the centre of mass of those heatmaps. These 2D poses are then normalized and scaled before lifting to 3D. 0.0001, the Adam optimizer uses a learning rate to train the lifting network. The generated 3D poses are also normalized before being passed to the detection network. As the different video samples contain diverse frames, all samples have been extended to 300 by adding null frames to maintain uniformity. The fall detection network has been trained with Cross Entropy Loss with Adam Optimizer taking the same learning rate with exponential decay. This network has been trained on one Azure Nvidia V100 GPU for 50 epochs.

4.4 Performance Comparison Analysis

In this section, the performance evaluation and comparative analysis of the designed network have been covered. Firstly, the performance of the individual components has been discussed, and then the performance of the whole network has been analyzed.

4.4.1 Performance of Compressed Video Streams

The video compression network has been designed and implemented in incremental order. The reconstructed images' qualitative performance, i.e., visual quality, is measured using Structural Similarity Index Measure (SSIM) and Peak Signal to Noise Ratio (PSNR). Flow End Point Error and TPF, Time Per Frame, measure the quantitative performance. A random emission step strategy is used to train the network. The number of emission steps is taken randomly from one to ten. The performance values of the compression network have been measured for each emission step, and their average has been given in Table 3.

Table 3: Average performance values

	Avg. SSIM	Avg. PSNR	Avg. EPE	Avg. TPF
Video compression network	0.9036	26.22	0.3199	0.02662

It is observed that the visual quality of the frame is improved with the addition of each emission. Flow End Point Error measures the error in motion information among subsequent frames, and Time Per Frame (TPF) represents the frame reconstruction time. The obtained values show that the error gradually decreases with each new emission step. The obtained values infer that including multiple emission steps improves the images' visual quality and reduces the flow error but with some nominal increment in frame reconstruction time. The detection quality from encoded frames will improve; after that, the activity recognition results will improve.

4.4.2 Pose Estimation

Some earlier works on pose estimation directly utilize 3D annotations of the chosen dataset. Still, we have used predicted poses to assess the performance and feasibility of the proposed system. Joint Detection Rate (JDR) indicates the pose estimation accuracy. The percentage of successful joint

detections is termed JDR. Any joint is considered a successful detection when the difference between ground truth and estimation is less than the threshold taken. The half distance between the head and neck is the threshold here.

Table 4 shows the JDR results of our network on some joints, including the wrist, elbow, spine, and head. The joints having more than 90% JDR are taken as accurate predictions. The JDR of the thumb and ankle is lower than the elbow and wrist due to inaccuracy in 2D poses or frequent occlusion. The same has also been reflected in the last row of the above table.

Table 4: Joint detection accuracy of some major joints

Joint type	JDR
Head	98.01
Base Spine (BS)	99.28
Left Elbow (LE)	98.15
Left Wrist (LW)	97.51
Right Elbow (RE)	94.17
Right Wrist (RW)	91.73
Right Ankle (RA)	71. 02

The skeleton information of the chosen data set contains 25 joints. However, to keep our system lightweight and less computative, we have also selected the aggregation of 16 joints for our convenience and computed the mean JDR for both joint structures. The mean JDR of 16 joints is greater than 25 points, as depicted in Fig. 9.

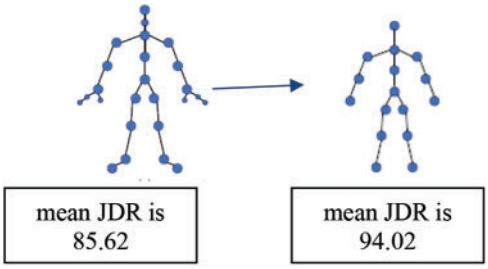

Figure 9: Skeleton structure and mean JDR of 25 joints and 16 joints as computed by our network

4.4.3 Activity Recognition

The performance of the proposed system has been assessed in terms of the accuracy of events detected. The proposed network achieves a fall accuracy of 99.21% and overall accuracy of 99.72% for all three classes. The comparative results with some other networks have been given in Table 5. The graphical comparison is shown in Figs. 10 and 11. The table also infers that earlier networks focused only on detection tasks, but we extended our network to monitor other events, i.e., rest and activity, efficiently.

Table 5: Accuracy measurements of the proposed system with some other networks

Methods	Network	Event	Input	Feature	Fall accuracy	Overall accuracy
Tsai et al. [24]	Conv (1D)	Fall	Depth	Pose	99.20%	99.20%
Shojaei-Hashemi et al. [21]	LSTM	Fall	Depth	Pose	96.12%	96.12%
Xu et al. [22]	Conv (2D)	Fall	RGB	Pose	91.70%	91.70%
Chen et al. [29]	Conv (1D)	Fall	RGB	Pose	99.83%	99.83%
Sykes [26]	DNN	Fall	RGB	Pose	92.59%	92.59%
Proposed system	Conv (1D)	Rest, activity, fall	RGB	Pose	99.63%	99.77%

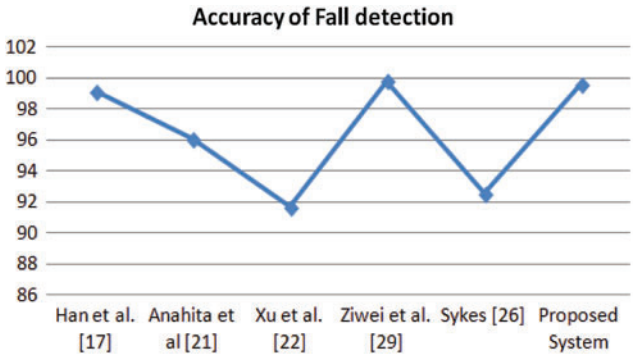

Figure 10: Accuracy comparison of fall detection of the proposed system with some state-of-art works

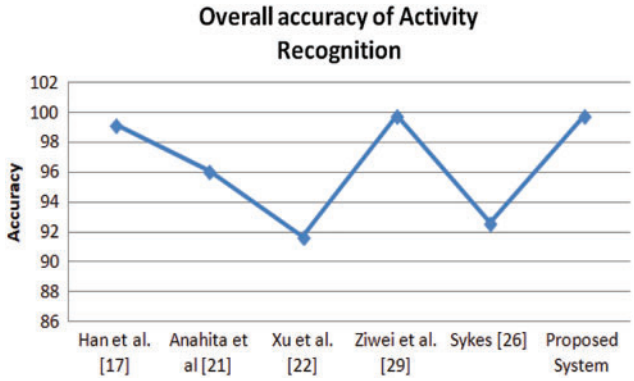

Figure 11: Accuracy comparison of activity recognition of the proposed system with some state-of-art works

4.4.4 Discussion

This work contributes towards making video-based patient monitoring more practical and less computation intensive. In addition, to fall detection, it provides patient activity and rest time monitoring, which can be utilized in routine to improve patient well-being. It also provides an approach to achieving high-accuracy fall detection with lower computation costs by sharing the network with

rest and activity detection tasks. Compared to earlier works, rather than scientific advancement in specific activity detection, we expand the current fall detection methods to more activity detection without increasing the computational cost and with detection accuracy comparable to existing works. This network may also provide accounting information by installing corresponding counters. This experimentation is carried out on the generalized dataset. Its performance can be further evaluated on varied and specific datasets in the extended version of this work. Moreover, using ResNet Mixer blocks in activity recognition networks classifies the given input efficiently with an accuracy of 99.77%, but several factors of real-time scenarios have not been considered. This work may be further extended for its more practicality.

4.4.5 Limitations

One of the significant limitations of this work is the lack of real-world evaluation. The dataset captures various individual actions and steps from different angles and viewpoints. This diversity provides some mitigation. Other associated challenges are complex background or lighting variation, which affects the accuracy, and must be addressed. Moreover, diverse situations that may lead to imminent falls and the impacts of various factors like loss function or input joints, which may affect the performance of these systems, may also be explored.

5 Conclusion

A lightweight compressed video stream-based patient activity monitoring system has been proposed in this paper. The main focus of this work is to improve the value of the existing network by expanding fall detection to activity recognition without impacting the accuracy of fall detection and increasing computational costs. The whole system is an integration of three sub-networks. A CNN-based video compression network, an LPN for human pose estimation, and a ResNet Mixer block-based activity recognition network are used collectively to accomplish the task.

The major limitation of this research work is the lack of real-world evaluation. The actions in the dataset are captured from different angles and viewpoints. This diversity provides some mitigation. Moreover, some challenges associated with such vision-based systems are complex background or lighting variation, which affects the accuracy, and need to be addressed. The diverse situations that may lead to imminent falls must be explored. The impacts of various factors like loss function or input joints, which may affect the performance of these systems, may also be studied in future works.

The system has been trained and tested over a standard dataset, resulting in an overall accuracy of approx. 99.77% with a detection accuracy of 99.63%. The performance analysis of the obtained results shows the efficient detection of different events by the proposed system, which may help monitor the improvements in inpatient falls. Future enhancements can include integrating activities such as detecting seizures or critical events, multi-camera setups, real-time alerts, data privacy measures, and validation in natural healthcare settings.

Acknowledgement: We are thankful to Majmaah University for its help in conducting this study.

Funding Statement: Sultan Alharby would like to thank the Deanship of Scientific Research at Majmaah University for funding this work under Project No. R-2023-667.

Author Contributions: Study conception and design: S. Yadav, P. Gulia; data collection: S. Yadav; analysis and interpretation of results: S. Yadav, N. S. Gill; draft manuscript preparation: S. Yadav, P. Gulia, N. S. Gill. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data will be made available on request from the submitting author.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. Martinez, D. Ahmedt-Aristizabal, T. Väth, C. Fookes and R. Stiefelhagen, "A vision-based system for breathing disorder identification: A deep learning perspective," *IEEE Engineering in Medicine and Biology Society*, vol. 1, pp. 6529–6532, 2019.

[2] D. Ahmedt-Aristizabal, S. Denman, K. Nguyen, S. Sridharan, S. Dionisio *et al.,* "Understanding patients' behaviour: Vision-based analysis of seizure disorders," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2583–2591, 2019.

[3] S. Liu, Y. Yin and S. Ostadabbas, "In-bed pose estimation: Deep learning with shallow dataset," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 7, pp. 1–12, 2019.

[4] V. B. Semwal, A. Gupta and P. Lalwani, "An optimized hybrid deep learning model using ensemble learning approach for human walking activities recognition," *The Journal of Supercomputing*, vol. 77, pp. 12256–12279, 2019.

[5] A. Masalha, N. Eichler, S. Raz, A. Toledano-Shubi, D. Niv *et al.,* "Predicting fall probability based on a validated balance scale," in *Proc. of IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, Seattle, WA, USA, pp. 302–303, 2020.

[6] W. Viriyavit and V. Sornlertlamvanich, "Bed position classification by a neural network and Bayesian network using non-invasive sensors for fall prevention," *Journal of Sensors*, vol. 2020, pp. 1–14, 2020.

[7] J. Gutierrez, V. Rodríguez and S. Martin, "A comprehensive review of vision-based fall detection systems," *Sensors*, vol. 21, no. 3, pp. 1–50, 2021.

[8] V. Bijalwan, V. B. Semwal and T. K. Mandal, "Fusion of multi-sensor-based biomechanical gait analysis using vision and wearable sensor," *IEEE Sensors Journal*, vol. 21, no. 13, pp. 14213–14220, 2021.

[9] R. Jain and V. B. Semwal, "A novel feature extraction method for pre-impact fall detection system using deep learning and wearable sensors," *IEEE Sensors Journal*, vol. 22, no. 23, pp. 22943–22951, 2022.

[10] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng *et al.,* "Deep, high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 3349–3364, 2021.

[11] Z. Cao, G. Hidalgo, T. Simon, S. Wei and Y. Sheikh, "Openpose: Real-time multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 172–186, 2021.

[12] S. Yang, Z. Quan, M. Nie and W. Yang, "Transpose: Keypoint localisation via transformer," in *Proc. of IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 11782–11792, 2021.

[13] M. Kocabas, N. Athanasiou and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *Proc. of IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 5252–5262, 2020.

[14] B. Wandt, M. Rudolph, P. Zell, H. Rhodin and B. Rosenhahn, "Canonpose: Self-supervised monocular 3D human pose estimation in the wild," in *Proc. of IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 13289–13299, 2021.

[15] C. Menacho and J. Ordoñez, "Fall detection based on CNN models implemented on a mobile robot," in *Proc. of 17th Int. Conf. on Ubiquitous Robots (UR)*, Kyoto, Japan, pp. 284–289, 2020.

[16] X. Wang, J. Ellul and G. Azzopardi, "Elderly fall detection systems: A literature survey," *Frontiers in Robotics and AI*, vol. 7, pp. 1–23, 2020.
[17] T. Tsai and C. Hsu, "Implementation of a fall detection system based on 3D skeleton for deep learning technique," *IEEE Access*, vol. 7, pp. 153049–153059, 2019.
[18] V. Bijalwan, V. B. Semwal and T. K. Mandal, "Fusion of multi-sensorbased biomechanical gait analysis using vision and wearable sensor," *IEEE Sensors Journal*, vol. 21, no. 13, pp. 14213–14220, 2021.
[19] P. Patil, K. S. Kumar, N. Gaud and V. B. Semwal, "Clinical human gait classification: Extreme learning machine approach," in *Proc. of Int. Conf. on Advances in Science, Engineering and Robotics Technology (ICASERT)*, Dhaka, Bangladesh, pp. 1–6, 2019.
[20] A. A. Salamai, N. Ajabnoor and A. M. Khawaji, "Deep learning model for early weed detection in agriculture application," *International Journal of Advances in Applied Computational Intelligence*, vol. 2, no. 1, pp. 23–28, 2022.
[21] A. Shojaei-Hashemi, P. Nasiopoulos, J. J. Little and M. T. Pourazad, "Video-based human fall detection in smart homes using deep learning," in *IEEE Int. Symp. on Circuits and Systems (ISCAS)*, Florence, Italy, pp. 1–5, 2018.
[22] Q. Xu, G. Huang, M. Yu and Y. Guo, "Fall prediction based on key points of human bones," *Physica A: Statistical Mechanics and its Applications*, vol. 540, pp. 123205–123237, 2020.
[23] N. Lu, Y. Wu, L. Feng and J. Song, "Deep learning for fall detection: Three-dimensional CNN combined with LSTM on video kinematic data," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 1, pp. 314–323, 2019.
[24] T. H. Tsai and C. W. Hsu, "Implementation of fall detection system based on 3D skeleton for deep learning technique," *IEEE Access*, vol. 7, pp. 153049–153059, 2019.
[25] C. Zhong, W. W. Y. Ng, S. Zhang, C. D. Nugent, C. Shewell *et al.,* "Multi-occupancy fall detection using non-invasive thermal vision sensor," *IEEE Sensors Journal*, vol. 21, no. 4, pp. 5377–5388, 2021.
[26] E. R. Sykes, "An analysis of current fall detection systems and the role of smart devices and machine learning in future systems," in *Advances in Information and Communication*, vol. 652, Cham: Springer, pp. 502–520, 2023.
[27] S. Mobsite, N. Alaoui, M. Boulmalf and M. Ghogho, "Semantic segmentation-based system for fall detection and post-fall posture classification," *Engineering Applications of Artificial Intelligence*, vol. 117, pp. 105616–105672, 2023.
[28] Y. Cheng, B. Yang, B. Wang and R. T. Tan, "3D human pose estimation using spatio-temporal networks with explicit occlusion training," in *Proc. of AAAI Conf. on Artificial Intelligence*, New York, USA, vol. 34, pp. 10631–10638, 2020.
[29] Z. Chen, Y. Wang and W. Yang, "Video-based fall detection using human poses," in *Proc. of CCF Conf. on Big Data*, Guangzhou, China, pp. 283–296, 2022.
[30] X. Ma, H. Wang, B. Xue, M. Zhou, B. Ji *et al.,* "Depth-based human fall detection via shape features and improved extreme learning machine," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 6, pp. 1915–1922, 2014.
[31] Z. Zhang, C. Conly and V. Athitsos, "Evaluating depth-based computer vision methods for falldetection under occlusions," in *Advances in Visual Computing*. Cham: Springer, pp. 196–207, 2014.
[32] K. Bogdan and M. Kepski, "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Computer Methods and Programs in Biomedicine*, vol. 117, no. 3, pp. 489–501, 2014.
[33] S. Gasparrini, E. Cippitelli, E. Gambi, S. Spinsante, J. Wåhslén *et al.,* "Proposal and experimental evaluation of fall detection solution based on wearable and depth data fusion," in *Suzana Loshkovska and Saso Koceski, ICT Innovations*. Cham: Springer, pp. 99–108, 2016.
[34] C. Liu, Y. Hu, Y. Li, S. Song and J. Liu, "PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding," in *Proc. of Workshop on Visual Analysis in Smart and Connected Communities*, California, USA, pp. 1976–1977, 2017.

[35] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Duan *et al.,* "NTU RGB + D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2684–2701, 2020.

[36] A. S. Aziz, H. K. Mohamed and A. Abdelhafeez, "Unveiling the power of convolutional networks: Applied computational intelligence for arrhythmia detection from ECG signals," *International Journal of Advances in Applied Computational Intelligence*, vol. 1, no. 2, pp. 63–72, 2022.

[37] K. Alakkari, M. Abotaleb, A. Badr, A. Kadi, A. M. Ghazi Al-Khatib *et al.,* "Modelling weather conditions using encoder-decoder and attention based on LSTM deep regression model," *International Journal of Advances in Applied Computational Intelligence*, vol. 1, no. 2, pp. 8–29, 2022.

[38] M. Ullah, M. M. Yamin, A. Mohammed, S. D. Khan, H. Ullah *et al.,* "Attendance-based LSTM network for action recognition in sports," in *Proc. of IS&T Int. Symp. on Electronic Imaging: Intelligent Robotics and Industrial Applications Using Computer Vision*, West Lafayette, USA, pp. 3021–3026, 2021.

[39] S. Yadav, P. Gulia and N. S. Gill, "Flow-MotionNet: A neural network based video compression architecture," *Multimedia Tools Applications*, vol. 81, pp. 42783–42804, 2022.

[40] A. Shahroudy, J. Liu, T. T. Ng and G. Wang, "NTU RGB + D: A large scale dataset for 3D human activity analysis," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 1010–1019, 2016.

[41] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona *et al.,* "Microsoft COCO: Common objects in context," in *Computer Vision–ECCV 2014*, Cham, Springer, pp. 740–755, 2014.