Human activity recognition and fall detection using video and inertial sensors

by

Zhanggir Yergaliyev

Submitted to the School of Engineering and Digital Sciences
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

May 2022

Author  Zhanggir Yergaliyev
School of Engineering and Digital Sciences
Apr 29, 2022

Certified by
Adnan Yazici
Department Chair of Computer Science
Thesis Supervisor

Accepted by
Vassilios D. Tourassis
Dean, School of Science and Technology

Human activity recognition and fall detection using video and inertial sensors

by

Zhanggir Yergaliyev

Submitted to the School of Engineering and Digital Sciences
on Apr 29, 2022, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

Falls are a crucial problem for elderly people. Early detection of falls may prevent or attenuate possible negative consequences for elderly people. There is a number of scientific articles on the topic of detecting falls using machine learning techniques. While some of them focus on fall detection systems based on scalar body sensors, others apply vision based detection. The goal of this thesis is to try to perform a fusion of inertial sensor based and video sensor based modules to provide a more robust solution, as each method has their own drawbacks in terms of both performance and feasibility.

Thesis Supervisor: Adnan Yazici
Title: Department Chair of Computer Science

Acknowledgments

I am deeply grateful to my primary supervisor, Adnan Yazici, who guided me throughout this project. Also, I wish to acknowledge the help provided by Enver Ever and Hakan Yatbaz, and the Department of Computer science of the Nazarbayev University.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the development of Internet of Things (IoT), our capability to retrieve data from varying environments for different industries have gone to a new level. Additionally, with the increasing capability of machine learning in different fields, now we are able to build an infrastructure which offers new opportunities to improve the quality of our lives, particularly in healthcare.

According to [13], the use of digital technologies to monitor and gather medical and other health information from patients and digitally transfer this data to healthcare professionals for evaluation is known as remote patient monitoring. When it comes to monitoring the heath of elderly population, one of the high-risk scenarios for them is falling, causing their body to hit brutally the ground [19]. Although a fall can happen to anyone, seniors are especially affected by such incidents. Falls are the most common cause of the injuries to the elderly, and they are also responsible for over 60% of significant injuries suffered by people of all ages according to population-based studies [19]. In their report, the World Health Organization (WHO) mentions that the percentage of falls for people over 70 years is estimated to be 32-42% [31]. This shows that older people need care and protection, especially when they have some health problems.

To provide the care and protection needed, the actions of the user should be identified first, which is one of the challenges for intelligent systems. Luckily, with the fast advancements in machine learning domain, it is possible now to utilise multitude

of sensors such as inertial sensors, cameras etc. to identify human activity recognition (HAR).

HAR is a problem which consists of a few general stages. As with any machine learning problem, HAR models need to be trained on some data. The data can be obtained in two ways: either by collecting data firsthand, or by downloading open-source datasets from the Internet. Next step is data preprocessing, since the raw data cannot be applied to the training process of machine learning model, unless it was preprocessed. Feature extraction, which is a method to shrink the number of resources required to describe a set of data, is done. And finally, the last step is to train a classification model. To recognize various activities and falls, HAR system is required to be able to understand regular human activities. Some of the popular classification methods for HAR are Random Forest, SVM, CNN, LSTM, and k-Nearest Neighbors [10].

In general, models designed to recognize activities performed by humans utilize data from two different sources: inertial sensors and video cameras. Inertial sensors may include accelerometer, which is an instrument used for measuring the acceleration of an object, gyroscope (measuring angular velocity), and magnetometer (measuring magnetic induction). Another widespread approach for HAR is video-based activity recognition. Video cameras, unlike scalar body sensors, provide a non-intrusive way to record data pertinent to activity recognition. Video data, on the other hand, necessitates more memory on the device [9]. There is an emerging trend for designing hybrid systems for action recognition in response to concerns that fall detection systems employing solely inertial sensors produce too many false alarms and video-based activity recognition systems lack accuracy and precision. Scalar sensor data and video data are commonly used in hybrid approaches for HAR and fall detection.

Therefore, the main goals of this study are:

- To be able to properly analyze and identify various activities performed by humans in home conditions.

- To build state-of-the-art vision-based activity recognition model incorporating

falls.

- To build state-of-the-art inertial sensor based activity recognition model incorporating falls.
- to build first multimodal activity recognition model which recognizes falls.

A lot of work have been done on activity recognition from either inertial or video sensors. There are only a few works that include both inertial and video sensors. In contrast, we will try to perform a fusion of inertial sensor based and video sensor based modules to produce a solution which incorporates two approaches and mitigates the drawbacks of each individual approach. This work proposes a convolutional neural network (CNN) architecture for recognizing activities from DMLSmartActions dataset. Additionally, a Transformer model is developed for the videos of UP-Fall dataset. Experiments show that both model reaches state-of-the-art accuracy scores for their tasks.

The rest of the paper is organized as follows. Section 2 introduces some relevant related works on this area of research. Section 3 describes methodologies used to achieve main goals of this study. The results of our activity recognition approach, and their evaluation are presented in Section 4. Finally, Section 5 concludes this thesis paper.

Chapter 2

Related work

2.1 HAR from inertial sensor data

As mentioned earlier, there are different ways in which HAR can be implemented. Activities performed by people can be measured by inertial sensors such as accelerometer, gyroscope, and magnetometer.

[24] presented a system which combines three wearable inertial sensors with a Frequency-Modulated Continuous Wave (FMCW) radar. In their experimental setup, inertial sensors were placed on the ankle, waist, and wrist of the subjects. Unlike in many other similar experimental setups, [24] collected the data so that transitions between various activities occur at random times. After collecting, preprocessing, and feature extraction steps, they have built a bi-LSTM (bidirectional Long Short-Term Memory) network to classify falls and other activities. The model has scored an accuracy of 96% on the data of the subjects unknown to the classifier.

Similarly, [3] developed a model which recognized activities of daily living and falls using accelerometer data. However, unlike [24], the authors have utilized public datasets for this purpose, namely SisFall and UMAFall datasets. The data from those datasets are not from independent inertial sensors, but from an accelerometer sensor built in smartphones. Authors proposed a novel ConvLSTM network, which utilizes the advantages of convolutions for feature extraction, and applies LSTM for feature processing. Using the proposed architecture, they perform binary classification of

activities as fall and non-fall. The proposed ConvLSTM model achieved 95% accuracy on UMAFall dataset and 98.39% on SisFall dataset, which makes it state-of-art for corresponding datasets.

Utilization of accelerometer built in smartphones became a desirable solution for human activity recognition, since it is practical and easy to implement. [1] also built a HAR model based on smartphone accelerometer data available in public. However, instead of implementing popular classification methods for this task such as CNN and LSTM, they offered an algorithm called Ameva, which is an original discretization, selection, and classification technique for HAR. The key feature of Ameva is low-consuming battery system. The algorithm attained the average accuracy of 95% on USC-HAD, WISDM and Shoaib datasets.

2.2 HAR from video data

Apart from inertial sensor data based HAR, the next most popular method for activity recognition is video based HAR. Unlike scalar body sensors, video cameras offer non-intrusive approach to record data relevant for activity recognition. But, on the other hand, video data demands larger amounts of memory on the device.

One of the first non-intrusive experimental setups to record activity data using video cameras was developed by [6]. They have developed a system to monitor activity of occupants in a smart home environment. The experimental setup for the dataset called DMLSmartActions consisted of two RGB cameras and one Kinect, including 18 subjects (12 male and 6 female). The classes represent the following distinct actions performed by subjects of the dataset: falling down, drinking, dropping and picking upsomething on the floor, picking up something, putting something, cleaning the table, reading, sitting down, standing up, using a cellphone, walking, writing. They have also built a classification algorithm for the data by implementing SVM, which was one of the state-of-art methods in 2014. The highest classification accuracy result was 58.2%. In 2019, [27] built a CNN model based on publicly available dataset published by [6]. The architecture of CNN built by [27] consisted of five convolutional

layers, four max pooling layers and three fully connected layers. The proposed CNN reached 82.41% accuracy rate on DMLSmartActions and became state-of-the-art for that dataset.

So far, all described systems focus on activity recognition of a single person. In 2020, [34] have published a paper which presents a real-time multiple-person activity recognition system based on deep learning. It is a complex system which captures a video stream of the scene, uses YOLOv3 to identify coordinates of bounding boxes of a person in the frame, implements FaceNet recognition approach to perform face recognition, and is able to execute automatic "zoom-in" when people in the frame are located too far from the camera. After all those steps, authors have developed a convolutional network-based architecture to classify activities, including falls. The model has scored 90.79% accuracy on NTU RGB+D dataset.

Similar to [34], [25] have implemented an action recognition and fall detection system by utilizing YOLOv3 to detect objects in the frame. They have combined YOLOv3 with LiteFlowNet, a light CNN for the estimation of optical flow, and called it YCL. YCL have attained an accuracy rate of 93.74% on the dataset collected and developed by the authors. Moreover, a smartphone application which sends an alert SMS to the next of kin of a subject once the fall action was confirmed was developed by the authors.

2.3 Hybrid HAR approaches

As a response to claims that fall detection systems using only inertial sensors produce too many false alarms and video based HAR systems lack accuracy and precision, some researchers started developing hybrid architectures for action recognition. Hybrid approach for HAR and fall detection typically employ both inertial sensor data and video data. In 2014, [21] created a multimodal dataset called UR Fall. The data was collected from an IMU and two Kinects. 5 subjects performed 70 sequences of activities of daily living and falls. Furthermore, subjects performed two types of falls: falling from chair and falling from a standing position. Authors devel-

oped a threshold-based fall detection technique, whereby a fall is firstly indicated by accelerometer, then the suggested model calculates features from inertial and video data, and executes SVM classifier to validate the fall.

In 2019, [26] produced a similar dataset called UP-Fall. UP-Fall dataset consists of 17 subjects performing 6 activities of daily living and 5 types of falls. To collect data, they used 3-axis accelerometer, 3-axis gyroscope, two cameras, electroencephalograph (EEG) headset to measure raw brainwave signal, and six infrared sensors for measuring fluctuations in interruption of the optical devices. UP-Fall contains 812 GB of data which stores 296,364 samples of images and raw sensor signals. Furthermore, authors tested the performance of several machine learning models such as Random Forest, SVM, Multi-Layer Perceptron, k-Nearest Neighbors on the proposed dataset. Random Forest classifier turned out be the model with highest accuracy score of 95.88%.

The most recent work that implements a hybrid approach to fall detection was presented by [23] in 2021. Authors propose a double-check method which is able to detect a fall of a person via IMU sensors and confirm the fall by analyzing images from RGB camera. As with previous methods, the IMU sensor if worn on the body of a subject. Since IMU generates sequential data, a recurrent neural network (RNN) was developed to classify falls from body worn sensor. When the RNN models outputs a fall action, the CNN model built on frames from RGB camera either confirms or discards the output of RNN. The RGB camera is mounted on a robot which is able to move across the house to get to a particular location and confirm a fall.

Chapter 3

Methodology

3.1 Video based activity recognition

3.1.1 Data description

Either inertial sensor data or video frames are required to train the human activity recognition model. Moreover, the list of activities performed must include falls. We utilize the data from three public dataset to that end.

Combined dataset for fall detection

The first dataset that we employ is compiled by combining video frames from several well-known fall-related datasets [2], [7], [8], [11], [33]. The dataset consists of 4 classes: sitting, standing, sleeping and falling down; containing 1500, 5900, 2300 and 2900 images respectively. The exemplary samples for each activity are depicted on Fig. 3-1.

DMLSmartActions dataset

The DMLSmartActions dataset from the University of British Columbia's Digital Multimedia Lab is also utilized to create the CNN model as an alternate dataset. In this dataset, individuals in a residential environment are monitored while conducting realistic activities. The dataset is publicly available and consists of one VGA stream

Figure 3-1: Some human activities from combined dataset

from a Kinect sensor and two HD streams from RGB cameras [6].

The model is trained and tested using the VGA stream. It includes 47 videos in which several different persons execute 6-12 different types of tasks. A script is developed to categorize the frames from each video into the appropriate classes. As a consequence, a total of 117,323 frames are obtained, divided into 12 classes. The classes present the following distinct activities carried out by the dataset's subjects: falling, drinking, dropping and picking up something on the floor, picking up, putting something, cleaning the table, reading, sitting down, standing up, using a smartphone, walking, and writing. Fig. 3-2 shows several examples of frames from the DMLSmartActions dataset.

UP-Fall dataset

UP fall [12] is a public dataset which was introduced in 2019. Seventeen teenage volunteers were engaged to help with several activities of daily living (ADL) and fall activities. The participants' average height and weight were 1.66 meters and 66.8 kg. The participants were between the ages of 18 and 24. The dataset includes six different ADLs as well as five different kinds of falls. In an experimental setup, various

Figure 3-2: Some human activities from DMLSmartActions

ambient sensors, wearable sensors, and cameras were used to record 11 activities. For the wearable sensors, an accelerometer, gyroscope, and electroencephalograph (EEG) data were provided. Five separate areas of the body were fitted with accelerometers and gyroscopes (neck, wrist, waist, ankle, and pocket). The recruited subjects' EEG sensors were put on their foreheads. Additionally, the experimental setup included two RGB cameras and 6 infrared sensors. For the video-based activity recognition module, we have only used image frames from one of the two RGB cameras. The experimental setup of for the dataset is depicted in Fig. 3-3.

Figure 3-3: The experimental setup from UP-Fall dataset

3.1.2 Proposed architectures

Due to the sensors utilised, video-based HAR systems often utilise computer-vision based solutions. Hence, the models in this section uses video samples collected from a standard RBG camera, web camera, or Kinect device that can also capture the depth of the frames for the vision-based activity recognition.

One of the disadvantages of vision-based HAR is the users' sense of privacy might be violated as image frames contain identifiable people and their behavior. Therefore, safety and security of such sensitive data is crucial [18]. Computational cost and memory requirement are the other challenges for vision-based HAR. Due to the nature of image data, such models require a large number of parameters for training. Deep learning architectures are typically trying to create a more complex hyperplane for the classification problem, to come up with a more robust solution, compared to traditional machine learning models. Therefore, they demand a lot of computational resources to achieve better results on image frames. Therefore, vision-based activity recognition has its drawbacks in terms of privacy, memory and computational power requirements [9].

To tackle the difficulties raised above, we suggest a setup in which the camera is triggered by events observed by the inertial sensors rather than running continuously [20]. When the inertial sensor-based activity recognition module detects an abnormality, the vision-based HAR is utilized simply for confirmation of the fall. The following is a general scenario:

- A user with an accelerometer sensor on his body lives in a home setting and carries out his regular activities. The data is continuously captured and used as an input to the activity recognition model based on inertial sensors.
- If an abnormality such as a fall is detected, the relevant camera turns on and begins collecting data, which is then input into a multimedia sensor-based HAR model. The model determines whether or not the fall is verified.
- The notification should be transmitted to the appropriate authorities if the fall is confirmed.

- The camera is turned off, but the accelerometer sensors continue to collect data and analyze the person's actions, in case if the fall is not confirmed.

Because video data is only captured when an user's safety is likely to be at risk, the given scenario helps to reduce issues with user's sense of privacy. Moreover, because the camera is not working all of the time, the memory constraint is reduced. This enables proposed system to work more efficiently and robustly than the system presented in the existing literature.

Convolutional Neural Network based Model

The development of deep learning based approaches led to significant improvements in terms of accurateness of machine learning models. In particular, convolutional neural networks (CNN) are considered one of the best performers in image data classification. A CNN is a specific architecture of artificial neural network that is designed to process data in a pixel format [22]. Any video is essentially a set of images moving (replacing each other) in a particular rate. Thus, utilizing the power of CNN in recognizing current user activity is suitable for our situation.

We chose to develop a convolutional neural network from scratch to fit the architecture of our CNN model to the datasets we employed. Figure 1 shows the architecture we constructed for our CNN model. The following six layers make up the structure:

- Layer 1: 32 2D convolutional filters of size (3x3) followed by 2D max-pooling operation and a ReLU activation function.

- Layer 2: 32 2D convolutional filters of size (3x3) followed by 2D max-pooling operation and a ReLU activation function.

- Layer 3: 32 2D convolutional filters of size (3x3) followed by 2D max-pooling operation and a ReLU activation function.

- Layer 4: A fully connected layer with 64 neurons.

- Layer 5: A fully connected layer with 64 neurons.

- Layer 6: 4 output units followed by a softmax activation function.

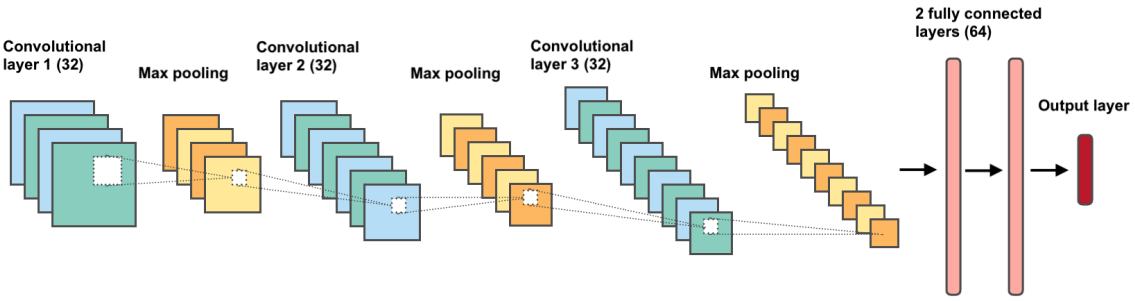

Figure 3-4: CNN model architecture for vision-based HAR

Hyperparameter optimization is applied in order to determine an optimal architecture for the model. TensorBoard is a tool that provides the statistics from the model's learning phase. Various models with different architectures are also investigated in addition to the hyperparameter tuning to find best performing model.

Transfer Learning

Another approach we used to train the activity recognition model is transfer learning. Transfer learning is a technology where the features pre-trained for a particular model are reused in a new machine learning model. Transfer learning allows better efficiency for model training [35]. We have utilized ResNet50 trained on ImageNet and added a dense layer with 128 nodes and ReLu activation function at the end. The model was trained on data from UP-Fall dataset's video frames.

Transformer

Another type of deep neural network architecture that is now gaining popularity in various challenging tasks is called transformer. A transformer weighs the significance of each component of the incoming data separately using the self-attention mechanism [14]. For the purpose of HAR, we have built a transformer model to classify.

To avoid overfitting, different data augmentations are applied. These are: horizontal flipping, random rotation, and random zooming.

The system takes in 72 by 72 image, transforms it into a grid of 6 by 6 patches, which will be inputted into our vision transformer model.

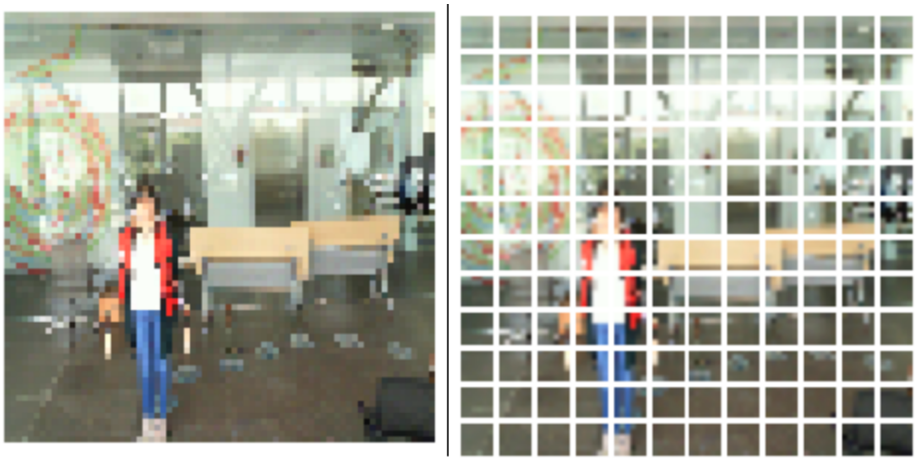

Figure 3-5: Image divided into patches for transformer classifier

The architecture of the vision transformer model is similar to the approach presented in [14]. The general algorithm is presented in Fig. 3-6.

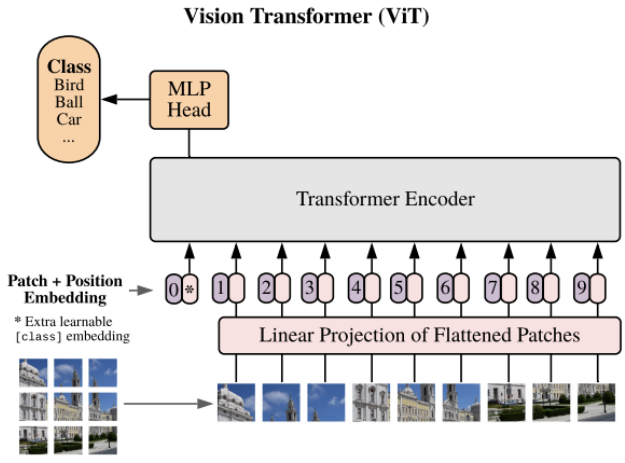

Figure 3-6: Vision transformer architecture [14]

3.2 Inertial sensor based activity recognition

3.2.1 Data description

As introduced in Section 3.1.1, the UP-Fall dataset is a hybrid dataset which contains data from both inertial sensors and camera. In this section, we utilize the inertial-sensor based portion of the dataset as the aim is to detect activities from inertial sensors.

For each wearable sensor in the dataset, 12 temporal and 6 frequency features were extracted (resulting in total of 756 features). The temporal features included:

- Mean
- Standard deviation
- Root mean square
- Maximal amplitude
- Minimal amplitude
- Median
- Number of zero-crossing
- Skewness
- Kurtosis
- First quartile
- Third quartile
- Autocorrelation

The frequency features are: mean frequency, median frequency, entropy, energy, principal frequency and spectral centroid.

3.2.2 Feature selection

For the feature selection, we followed a method described in [12]. A program takes the pre-selected features file and trains a Random Forest model with all features. After training, the model is fed features (adding one each time) and evaluates the results. Using the plots, we were able to decide how many of the pre-selected features are relevant to the process.

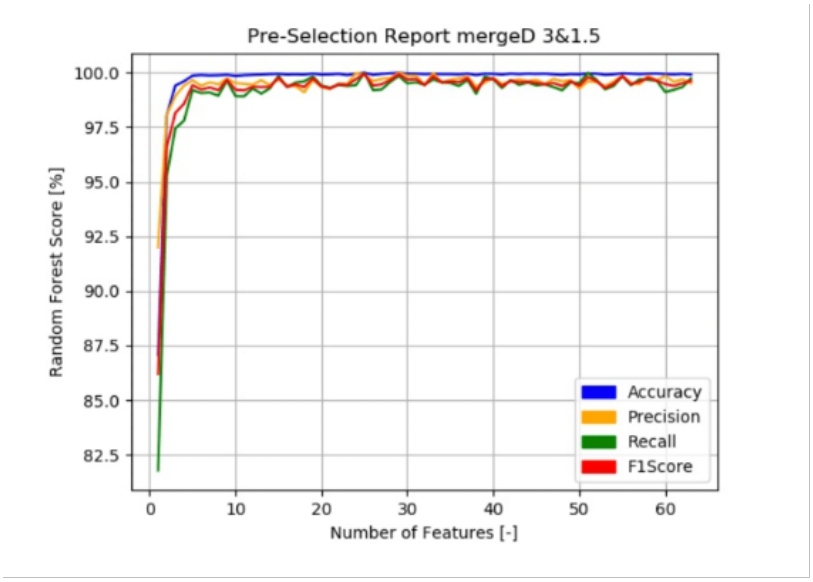

Figure 3-7: Feature selection process

3.2.3 Long short-term memory

Long short-term memory is one of the best performing networks when dealing time series type of data. Time series data is a set of data points seperated by successive equal time spaces. The data provided by [12] was homogenized with a sampling rate of 18 Hz. LSTM features feedback connections in contrast to typical feedforward neural networks. A recurrent neural network of this type may process complete data sequences instead of individual data points [16].

The first 90% of the data was taken for training and the last 10% for testing. We built sequences of 60 data points for our LSTM model. The architecture of the LSTM model consists of:

- Layer 1: LSTM layer with 128 units
- Layer 2: Dropout layer with a rate of 20%
- Layer 3: Batch normalization layer
- Layer 4: LSTM layer with 128 units
- Layer 5: Dropout layer with a rate of 20%

- Layer 6: Batch normalization layer
- Layer 7: LSTM layer with 128 units
- Layer 8: Dropout layer with a rate of 20%
- Layer 9: Batch normalization layer
- Layer 10: A fully connected layer with 32 neurons
- Layer 11: Dropout layer with a rate of 20%
- Layer 12: 11 output units followed by a softmax activation function

As with our CNN model for vison based activity recognition, we utilized Tensorflow to implement the LSTM model.

3.3 Hybrid approach for human activity recognition

3.3.1 Data description

The dataset which allowed us to train a hybrid system for human activity recognition, UP-Fall, has been introduced in Section 3.1.1. Since our goal is to build activity recognition system which learns from both inertial sensor and vision-based data, we have utilized data from accelerometers and gyroscopes located on five different parts of the subjects' bodies, and videos from an RGB camera for the proposed hybrid model. The dataset also contains recordings from EEG headset and six infrared sensors for measuring fluctuations in interruption of the optical devices.

3.3.2 Preprocessing

For the inertial sensor module, the process of preprocessing stays the same as described in Sections 3.2.1 and 3.2.2.

To be able to match the timestamps of inertial module data and recordings from camera, the preprocessing method for vision-based model is changed to optical-flow

method. This method makes it possible to determine the apparent displacements of objects in a picture series. These displacements can provide information on the correspondence between pixels in subsequent photos [30]. The optical flow was calculated by the method of the Horn and Schunck, described in [17]. After the processing the resulting dataset contains 800 features from the two cameras was produced [12].

3.3.3 Feature level fusion

To implement a multimodal activity recognition model, we propose the flow demonstrates in Fig 3-8.

We start with matching the data points derived from optical flow method with the timestamps from wearable sensor data. The wearable sensor data is preprocessed as described in Section 3.3.2.

For the vision based module, we built sequences of 60 data points. We chose to build a ConvLSTM model, because it is better than LSTM in handling spatiotemporal correlations [32].

After features from both LSTM and ConvLSTM are extracted, we concatenate them and define a new model. The concatenated features are trained on multilayer perceptron of the following architecture:

- Layer 1: A fully connected layer with 64 nodes
- Layer 2: Dropout layer with a rate of 50%
- Layer 3: Batch normalization layer
- Layer 4: A fully connected layer with 128 nodes
- Layer 5: Dropout layer with a rate of 50%
- Layer 6: Batch normalization layer
- Layer 7: A fully connected layer with 64 nodes
- Layer 8: Dropout layer with a rate of 50%

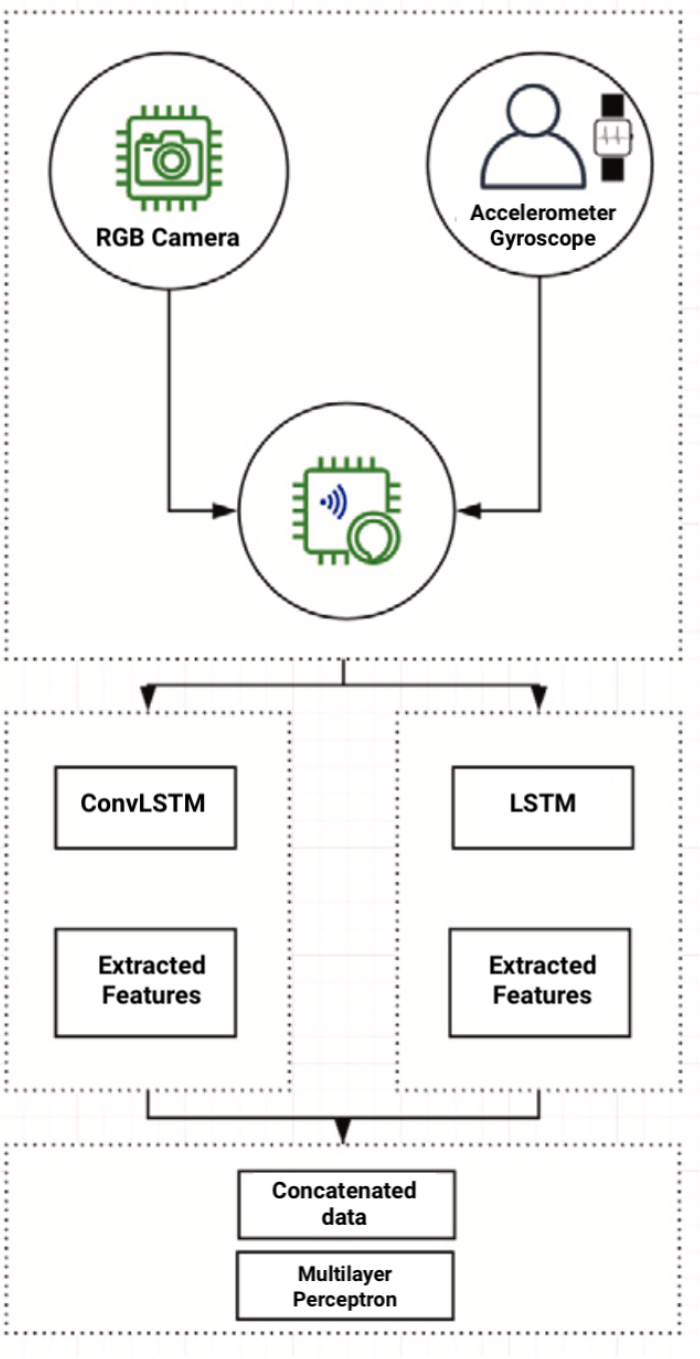

Figure 3-8: Feature level fusion

- Layer 9: 11 output units followed by a softmax activation function

The data is divided into 90% for training and 10% for testing.

Chapter 4

Results and evaluation

4.1 Vision based module

4.1.1 Combined dataset for fall detection

For the combined dataset for fall detection, the hyperparameters tuned by TensorBoard are the number of convolutional layers (ranging from one to three), the sizes of convolutional layers (32, 64, 128), and the number of dense layers (ranging from zero to two). After analyzing models with TensorBoard, we came up with the final model architecture which is illustrated in the Fig 2.

The class weights are assigned to each class of images because of the imbalanced nature of the dataset. The assignment is performed based on the number of samples of each class. For example, the "Sitting" class is generally misclassified since the number of samples in "Sitting" is approximately four times less than the number of samples in "Standing". To validate the model proposed, 5-fold cross-validation is performed. We have produced confusion matrices and classification reports with precision, recall, and f-1 scores for each fold. Fig. 3 depicts the confusion matrix calculated as an average across 5 folds. The average accuracy score obtained is 92.86%.

An alternative technique that can be used to balance the data sets is the hybrid approach that directly changes the number of samples in each class. The number of images in "fallen" and "sleeping" classes are similar enough and close to the average

Table 4.1: Comparison CNN architectures for fall dataset

Measure	Balanced CNN (class weights)	Balanced CNN (hybrid)
Average accuracy	92.85%	92.68%
Average loss	0.2092	0.2121

across all classes. Therefore, "sitting" is the main class that requires expanding and the "standing" class requires reduction. The former task was performed by increasing the number of samples in "sitting" by a factor of two by flipping each image horizontally. Following this, the number of samples in "standing" is cut down by half. As a result, the modified dataset ended up having 2,300-3,000 images for each class. The results with this dataset are similar to the results of the dataset obtained using the class weights. The average accuracy across 5-fold cross validations is roughly 92.69%. Fig. 4-2 represents the average confusion matrix of the model for the case where the hybrid approach is used to balance the dataset.

Table 4.1 shows the results for average accuracy and loss across 5-fold cross-validation. Results are presented for the dataset balanced through class weights as well as the dataset balanced with the hybrid approach.

4.1.2 DMLSmartActions dataset

The DMLSmartActions dataset of the Digital Multimedia Lab of the University of British Columbia is also used as an alternative dataset to develop the CNN model. This dataset contains monitoring of occupants in a home environment while they are performing realistic actions. The dataset consists of one VGA stream from a Kinect sensor and two HD streams from HD cameras [6], and is publicly available.

The VGA stream is used for training and testing of the model. It consists of 47 different videos, where several distinct people perform 6-12 different types of actions. A script is used to organize frames from each video according to corresponding classes. As a result, a total of 117,323 frames distributed across 12 classes are obtained. The classes represent the following distinct actions performed by subjects of the dataset:

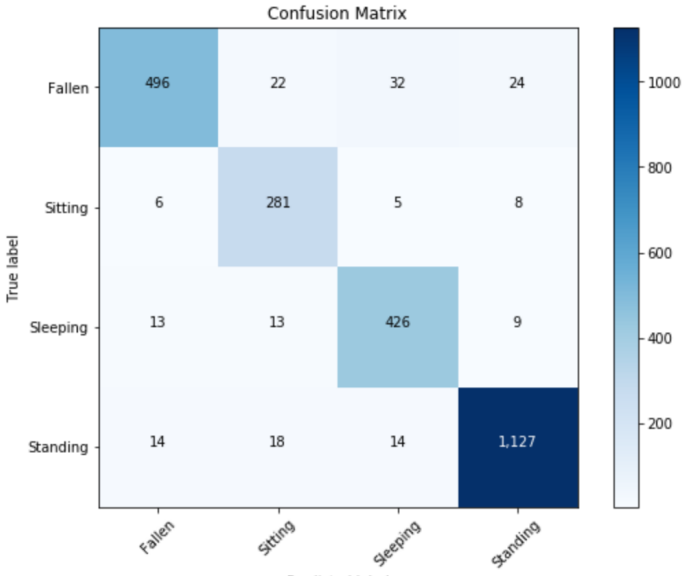

Figure 4-1: Confusion Matrix for Weighting Approach

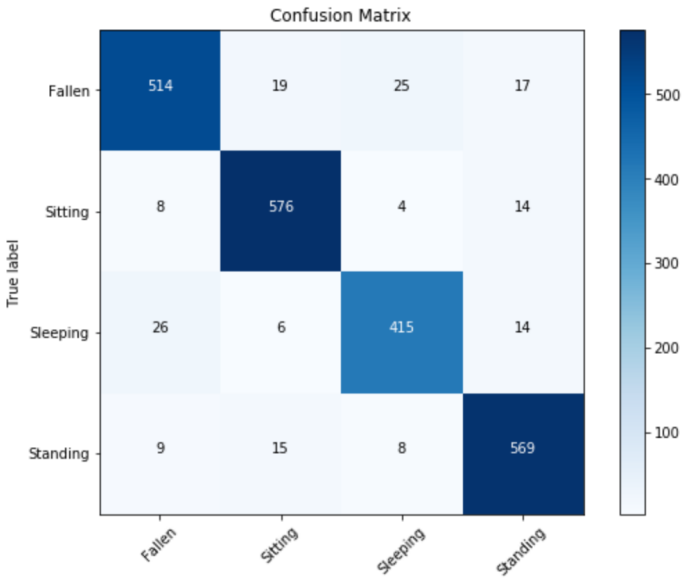

Figure 4-2: Confusion Matrix for Hybrid Approach

falling down, drinking, dropping and picking up something on the floor, picking up something, putting something, cleaning the table, reading, sitting down, standing up, using a cellphone, walking, writing. Some examples of frames from DMLSmartActions dataset are presented in Fig. 6.

Since the DMLSmartActions dataset is substantially larger than the combined dataset for fall detection presented in the previous subsection, the CNN architecture should be adjusted to fit a larger dataset. numerous model architectures are analyzed with varying hyper parameters to select the optimal one. The final architecture of the CNN model slightly differs from the original architecture presented in the Fig. 7. Layers 4 and 5, which consist of fully connected layers with 64 neurons, are replaced with one fully connected layer with 256 neurons, and the output has 12 units corresponding to 12 classes, instead of 4. We have assigned class weights to each class to handle the imbalanced nature of the dataset here as well. This time, we have performed 10-fold cross-validation, since the larger size of the dataset gives more samples for testing, hence allowing to train on a greater number of folds. The accuracy results across all 10 folds and their average score are presented in Table 4.2.

The results show that the model is able to classify most of the actions with accuracy of around 87%. Fig. 4-3 represents the confusion matrix for one of the folds of the proposed CNN model for DMLSmartActions dataset.

Most misclassifications occur with "walk" and "nothing" classes, which is understandable since "walk" contains many samples in which subjects are walking out of the frame, and only small parts of their bodies are seen, making it look like nobody is in the frame.

Various studies have been conducted on DMLSmartActions dataset in the literature, one of which is [27], which also uses a CNN with the 12 different classes used in this study with a total accuracy of 82.41%. Earlier studies such as [6], [4], and [5] also use the same dataset and traditional machine learning methods. All of these studies make use of SVM as their classification algorithm, and [5] reaches the highest accuracy rate with 79.9% for their proposed kernel method, which combines SVM with NNSC. Table4.3 provides a summary of the existing studies in the literature which

Table 4.2: Accuracy scores of CNN for DMLSmartActions

Fold	Accuracy	Loss
1	86.77%	0.2954
2	86.24%	0.3081
3	87.29%	0.3049
4	87.62%	0.2689
5	87.02%	0.2921
6	87.43%	0.2888
7	87.02%	0.3180
8	87.16%	0.2979
9	87.20%	0.2891
10	85.96%	0.3275
Average	86.97%	0.2991

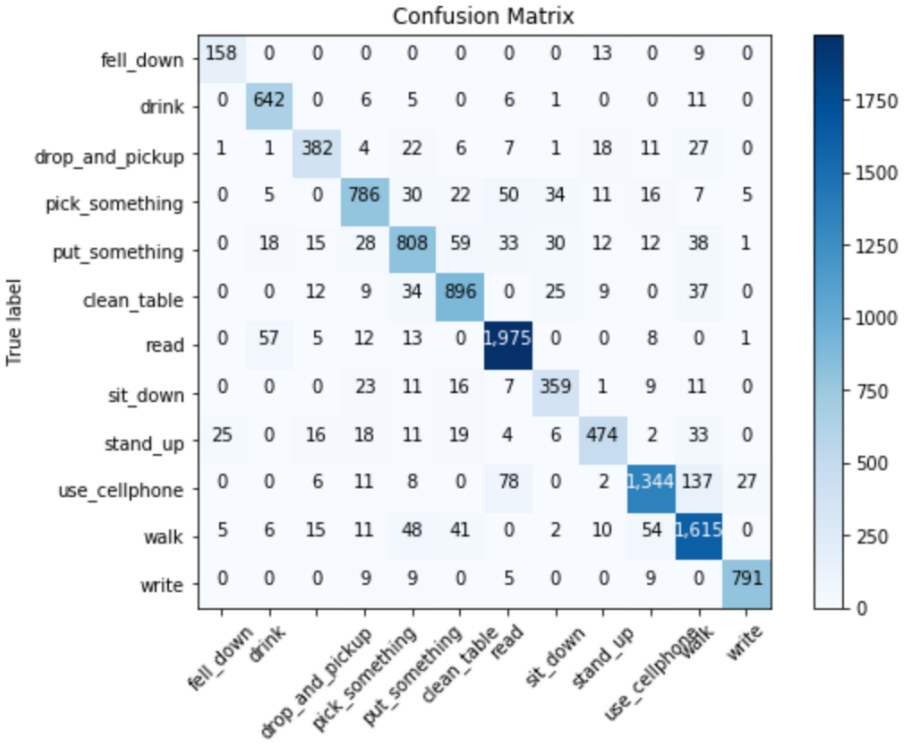

Figure 4-3: Confusion matrix for CNN for DMLSmartActions

makes use of multimedia information for activity recognition. The results in Table 7 clearly show that our models are superior to the existing ones in terms of accuracy for DMLSmartActions dataset. For the combined data set the overall accuracy of 92.86%.

Table 4.3: Comparison of the Proposed Model with the Literature

Ref	Model	Accuracy	Dataset
[6]	SVM using SC	58.20%	DMLSmartActions
[4]	SVM using NNSC	59.65%	DMLSmartActions
[5]	SVM with NNSC and proposed kernel	79.90%	DMLSmartActions
[27]	CNN	82.41%	DMLSmartActions
Proposed	CNN (Class Weighting)	92.86%	Combination of [1, 62, 9, 13, 8]
Proposed	CNN	86.97%	DMLSmartActions

4.1.3 UP-Fall dataset

The image frames from the videos of UP-Fall dataset are similar to DMLSmartActions dataset in terms of number of frames, number of classed, and image format. Thus, we decided to train the model with the same architecture from DMLSmartActions model with the images from UP-Fall. As a result, we came up with a classifier that has scored an accuracy of 98.55% on 5 folds, as represented on Table4.4.

Alternatively, we implemented a transfer learning model with ResNet50, as mentioned in Section 3.1.2. The transfer learning model has scored an accuracy of 99.7%.

Table 4.4: Performance of CNN model on UP-fall dataset's videos

Fold	Accuracy	Loss
1	99.13%	0.0377
2	98.05%	0.0689
3	97.88%	0.0686
4	98.95%	0.0380
5	98.77%	0.0405
Average	98.55%	0.0508

Table 4.5: Performance of different deep learning models on UP-fall videos

Model	Accuracy
CNN from scratch	98.55%
Transfer learning	99.70%
Transformer	99.87%

Table 4.6: Accuracy scores of LSTM for inertial data module

Fold	Accuracy
1	93.975%
2	91.121%
3	94.080%
4	91.966%
5	96.829%
6	90.169%
7	90.486%
8	89.641%
9	92.178%
10	96.829%
Average	92.727%

In the Transformer model, we have used AdamW optimizer, which an Adam optimizer with weight decay. It is an optimizer built into the Tensorflow addons library. The model was trained for 100 epochs and with 256 batch size. The Transformer model reached an accuracy of 99.87%, which the highest of the three approaches.

Since transformers gained such popularity in image classification recently, it is not surprising that it outperformed CNN and became our best model.

4.2 Inertial sensor based module

The performance of the LSTM model for wearable sensor based activity recognition is summarized in Table 4.6. The model was evaluated on a 10-fold cross-validation.

We have used 'Adam' as the optimizer for LSTM model. The model was trained for 25 epochs with a batch size of 64. Since LSTM is considered state-of-art in time

Table 4.7: Comparison of proposed model with other works on UP-Fall dataset

Ref	Modality	Classification Type	Model	Accuracy
[15]	vision-based	binary	CNN	95.64%
[28]	inertial	binary	LSTM	93.17%
[29]	inertial	binary	SVM, LR, DT, RF, KNN, NB	96%-99%
proposed	vision-based	multi class	CNN	98.55%
proposed	vision-based	multi class	Transformer	99.87%
proposed	vision-based	multi class	Transfer learning, ResNet50	99.7%
proposed	multimodal	multi class	LSTM + ConvLSTM	85.84%

series data classification, it is not surprising that the model scored a high accuracy.

4.3 Multimodal system

For the multimodal activity recognition model, we have scored an average accuracy of 85.835%, which is a good performance considering the task is multi-class activity classification (11 labels) rather than binary. Table 4.7 sums up the performances of proposed models, along with other works on the same dataset.

Several research on the UP-Fall dataset have been published in the literature. [15] offers a fall detection system based on a 2D CNN inference algorithm and several cameras, concentrating only on a vision-based approach. This method examines images within predetermined time periods and extracts features by employing an optical flow technique. The results showed that their proposed camera-based approach detects human falls and achieves an accuracy of 95.64% on binary classification. Similarly, [28] produced an LSTM model to detect falls using data from accelerometer and gyroscope only, and achieved accuracy of 93.17%. In 2021, [29] presented a work that focuses on three multimodal dataset, one of which is UP-Fall. However, they only use accelerometer data and perform a binary classification on whether fall action has occurred or not. They incorporate machile learning classifiers such as Random

Forest, Logistic Regression, Decision Trees, Naive Bayes, k-Nearest Neighbours and Support Vector Machines. Logistic Regression has achieved highest accuracy of 99%.

Our models, on the other hand, performs classification among all 11 classes, and contains the only multimodal activity recognition modal trained on UP-Fall.

However, we recognize that there is a problem in most of the works about activity recognition, either from inertial or video data. To evaluate the model properly, the out-of-sample data need to be separated correctly. The data cannot be shuffled before being split into train and test sets. Because in that case, the out-of-sample samples from test set would all have very close examples in the train set. Hence, it would be relatively easy for a model, as it overfits on the in-sample data, also overfit on test data. Instead, the data should be preprocessed in a way that a chunk of the sequences for testing is separated away before shuffling the data. Moreover, for temporal and time series data, out-of-sample needs to be a chunk of data in the future. That is why when preprocessing our sequences, we separated out last 10% of the data for testing. We claim that it is the most realistic way to do out-of-sample testing. As a result, our multimodal activity recognition model scored an accuracy of 85.84% on multiclass classification (11 labels) with a proper preprocessing method. The model fuses features obtained from LSTM for inertial data and ConvLSTM for vision-based approach.

Chapter 5

Conclusion

Falls are high-risk phenomenon which can cause a significant damage to a person, especially an elderly person. Since falls can lead to an emergency, building applications for successful detection of falls can make significant improvement to the lives of elderly population. To recognize falls, a system would need to distinguish them from other activities of daily living. Hence, fall detection and human activity recognition are inseparable parts of a system which is designed to help to prevent severe consequences from human falls.

A great deal of research has gone into activity recognition using inertial or video sensors. Only a few works incorporate both inertial and visual sensors. In this study, we have created a CNN model for recognizing actions from the DMLSmartActions dataset. Additionally, for the video module of the UP-Fall dataset, we constructed a Transformers model. Both models achieve state-of-the-art accuracy score. Additionally, data fusion is also applied to provide a multimodal intelligent HAR system in this thesis.

Bibliography

[1] Mobile activity recognition and fall detection system for elderly people using ameva algorithm. *Pervasive Mob. Comput.*, 34(C):3–13, January 2017.

[2] Kripesh Adhikari, Hamid Bouchachia, and Hammadi Nait-Charif. Activity recognition for indoor fall detection using convolutional neural network. In *Fifteenth IAPR International Conference on Machine Vision Applications, MVA 2017, Nagoya, Japan, May 8-12, 2017*, pages 81–84. IEEE, 2017.

[3] Mohamed Ilyes Amara, Abderrahmane Akkouche, Elhocine Boutellaa, and Hakim Tayakout. A smartphone application for fall detection using accelerometer and convlstm network. In *2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*, pages 92–96, 2021.

[4] S. Mohsen Amiri, Mahsa Pourazad, and Panos Nasiopoulos. Improved human action recognition in a smart home environment setting. *IRBM*, 35, 11 2014.

[5] S. Mohsen Amiri, Mahsa T. Pourazad, Panos Nasiopoulos, and Victor C. M. Leung. A similarity measure for analyzing human activities using human-object interaction context. In *2014 IEEE International Conference on Image Processing, ICIP 2014, Paris, France, October 27-30, 2014*, pages 2368–2372. IEEE, 2014.

[6] S. Mohsen Amiri, Mahsa T. Pourazad, Panos Nasiopoulos, and Victor C.M. Leung. Non-intrusive human activity monitoring in a smart home environment. In *2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013)*, pages 606–610, 2013.

[7] Morris Antonello, Marco Carraro, Marco Pierobon, and Emanuele Menegatti. Fast and robust detection of fallen people from a mobile robot. In *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*. IEEE, 2017.

[8] E. Auvinet, C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau. Multiple cameras fall dataset. In *Technical report 1350, DIRO - Université de Montréale*, 2010.

[9] Djamila Romaissa Beddiar, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. Vision-based human activity recognition: a survey. *Multim. Tools Appl.*, 79(41-42):30509–30555, 2020.

[10] Damien Bouchabou, Sao Mai Nguyen, Christophe Lohr, Benoit Leduc, and Ioannis Kanellos. A survey of human activity recognition in smart homes based on iot sensors algorithms: Taxonomies, challenges, and opportunities with deep learning. *CoRR*, abs/2111.04418, 2021.

[11] Imen Charfi, Johel Miteran, Julien Dubois, Mohamed Atri, and Rached Tourki. Optimized spatio-temporal descriptors for real-time fall detection: comparison of support vector machine and Adaboost-based classification. *Journal of Electronic Imaging*, 22(4):1 – 18, 2013.

[12] María de Lourdes Martínez-Villaseñor, Hiram Ponce, Jorge Brieva, Ernesto Moya-Albor, José Núñez-Martínez, and Carlos Peñafort-Asturiano. Up-fall detection dataset: A multimodal approach. *Sensors*, 19(9):1988, 2019.

[13] Pedro Dias, Miguel Cardoso, Federico Guede Fernández, Ana Martins, and Ana Londral. Remote patient monitoring systems based on conversational agents for health data collection. In Nathalie Bier, Ana L. N. Fred, and Hugo Gamboa, editors, *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2022, Volume 5: HEALTHINF, Online Streaming, February 9-11, 2022*, pages 812–820. SCITEPRESS, 2022.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[15] Ricardo Espinosa, Hiram Ponce, Sebastián Gutiérrez, Lourdes Martínez-Villaseñor, Jorge Brieva, and Ernesto Moya-Albor. A vision-based approach for fall detection using multiple cameras and convolutional neural networks: A case study using the up-fall detection dataset. *Computers in Biology and Medicine*, 115:103520, 2019.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[17] Berthold K.P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1):185–203, 1981.

[18] Im Jung. A review of privacy-preserving human and human activity recognition. *International Journal on Smart Sensing and Intelligent Systems*, 13:1–13, 01 2020.

[19] Pekka Kannus, Harri Sievänen, Mika Palvanen, Teppo Järvinen, and Jari Parkkari. Prevention of falls and consequent injuries in elderly people. *The Lancet*, 366(9500):1885–1893, 2005.

[20] Burak Kizilkaya, Enver Ever, Hakan Yekta Yatbaz, and Adnan Yazici. An effective forest fire detection framework using heterogeneous wireless multimedia sensor networks. *ACM Trans. Multimedia Comput. Commun. Appl.*, 18(2), feb 2022.

[21] Bogdan Kwolek and Michal Kepski. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Computer Methods and Programs in Biomedicine*, 117:489–501, 10 2014.

[22] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In David A. Forsyth, Joseph L. Mundy, Vito Di Gesù, and Roberto Cipolla, editors, *Shape, Contour and Grouping in Computer Vision*, volume 1681 of *Lecture Notes in Computer Science*, page 319. Springer, 1999.

[23] Deok-Won Lee, Kooksung Jun, Khawar Naheem, and Mun Sang Kim. Deep neural network–based double-check method for fall detection using imu-l sensor and rgb camera data. *IEEE Access*, 9:48064–48079, 2021.

[24] Haobo Li, Aman Shrestha, Hadi Heidari, Julien Le Kernec, and Francesco Fioranelli. Bi-lstm network for multimodal continuous human activity recognition and fall detection. *IEEE Sensors Journal*, 20(3):1191–1201, 2020.

[25] Xiaojie Lv, Zongliang Gao, Changshun Yuan, Meng Li, and Chao Chen. Hybrid real-time fall detection system based on deep learning and multi-sensor fusion. In *2020 6th International Conference on Big Data and Information Analytics (BigDIA)*, pages 386–391, 2020.

[26] Lourdes Martinez-Villaseñor, Hiram Ponce, Jorge Brieva, Ernesto Moya-Albor, José Nuñez Martinez, and Carlos Peñafort-Asturiano. Up-fall detection dataset: A multimodal approach. *Sensors*, 19:1988, 04 2019.

[27] Homay Danaei Mehr and Huseyin Polat. Human activity recognition in smart home with deep learning approach. In *2019 7th International Istanbul Smart Grids and Cities Congress and Fair (ICSG)*, pages 149–153, 2019.

[28] Md Jaber Nahian, Tapotosh Ghosh, Mohammed Uddin, Md. Maynul Islam, Mufti Mahmud, and M. Shamim Kaiser. Towards artificial intelligence driven emotion aware fall monitoring framework suitable for elderly people with neurological disorder. pages 275–286, 09 2020.

[29] Md. Jaber Al Nahian, Tapotosh Ghosh, Md. Hasan Al Banna, Mohammed A. Aseeri, Mohammed Nasir Uddin, Muhammad Raisuddin Ahmed, Mufti Mahmud, and M. Shamim Kaiser. Towards an accelerometer-based elderly fall detection system using cross-disciplinary time series features. *IEEE Access*, 9:39413–39431, 2021.

[30] Kenneth Norberg. *Audio Visual Communication Review*, 1(3):190–194, 1953.

[31] Seong-Hi Park. Tools for assessing fall risk in the elderly: a systematic review and meta-analysis. *Aging clinical and experimental research*, 30(1):1–16, 2018.

[32] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 802–810, 2015.

[33] S. Singh, S. A. Velastin, and H. Ragheb. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 48–55, 2010.

[34] Jen-Kai Tsai, Chen-Chien Hsu, Wei-Yen Wang, and Shao-Kang Huang. Deep learning-based real-time multiple-person action recognition system. *Sensors*, 20:4758, 08 2020.

[35] Yan Yan, Tianzheng Liao, Jinjin Zhao, Jiahong Wang, Liang Ma, Wei Lv, Jing Xiong, and Lei Wang. Deep transfer learning with graph neural network for sensor-based human activity recognition. *CoRR*, abs/2203.07910, 2022.