2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)
Bari, Italy. October 6-9, 2019

Deep Learning for Multimodal Fall Detection

Lourdes Martínez-Villaseñor, Hiram Ponce, Karina Pérez-Daniel
Universidad Panamericana. Facultad de Ingeniería.
Augusto Rodin 498, Ciudad de México, 03920, México
{lmartine,hponce,kperezd}@up.edu.mx

***Abstract*—Fall detection systems can help providing quick assistance of the person diminishing the severity of the consequences of a fall. Real-time fall detection is important to decrease fear and time that a person remains laying on the floor after falling. In recent years, multimodal fall detection approaches are developed in order to gain more precision and robustness. In this work, we propose a multimodal fall detection system based on wearable sensors, ambient sensors and vision devices. We used long short-term memory networks (LSTM) and convolutional neural networks (CNN) for our analysis given that they are able to extract features from raw data, and are well suited for real-time detection. To test our proposal, we built a public multimodal dataset for fall detection. After experimentation, our proposed method reached 96.4% in accuracy, and it represented an improvement in precision, recall and F$_1$-score over using single LSTM or CNN networks for fall detection.**

***Index Terms*—Fall detection, multimodal data, real-time system, deep learning, long short-term memory networks, convolutional neural networks**

I. Introduction

Falling is one of the most important health problems mainly among elderly people. “An estimated 646 000 fatal falls occur each year, making it the second leading cause of unintentional injury death, after road traffic injuries. Over 80% of fall-related fatalities occur in low- and middle-income countries, with regions of the Western Pacific and South East Asia accounting for 60% of these deaths. In all regions of the world, death rates are highest among adults over the age of 60 years” [1].

Fast detection and alarm, when a fall event has occurred, can help providing quick assistance of the person diminishing severe consequences of falling. Opportune fall detection decreases fear and time that a person remains laying on the floor after falling, which are important factors that determine the severity of the fall. For this reason, real-time detection is essential in a fall detection system.

Fall detection approaches are based on wearable sensors and/or context-aware systems. Context-aware systems use sensors deployed in the environment like cameras, floor sensors, infrared sensors, thermal sensors, pressure sensors, radar and microphones among others [2]. Wearable sensors are based on accelerometer, gyroscope and other sensors in different devices or more recently embedded in smart phones.

In recent survey of fall detection systems, Xu et al. [3] , regarding sensors, wearable sensors based in accelerometers and Kinect are the most recent trends for fall detection. Nevertheless multimodal approaches are appearing in order to gain more precision and robustness. Regarding the algorithms used for fall detection, the authors study reveals that the tendency of algorithms used has shifted from conventional algorithms, namely threshold-based, rule-based and shape-based, to machine learning techniques. In early works, authors claim that threshold is adequate for real-time processing [4]. Since new advanced sensors perceive more detail of human activities, threshold method became inadequate to achieve this goal [3]. Machine learning methods have gained popularity in fall detection systems. The most cited papers have not adopted deep learning. Nevertheless, it is recently used mainly in context-based approaches.

On the other hand, a main challenge of all fall detection systems is to reduce false positives. One way to address this problem is using multiple fall detection modalities combined [5]. Perry et al. [6] suggested in their evaluation of real-time fall detection approaches that the combination of other sensors with accelerometers provides a more robust and reliable detector.

In this paper, we propose a multimodal fall detection system based in wearable sensors, ambient sensors and vision devices. We build a multimodal dataset called UP-Fall Detection that includes 12 activities performed by 17 subjects based on information collected from five inertial measurement units as wearable sensors, one electroencephalograph (EEG) headset, six infrared sensors as a grid, and two cameras. Deep learning technologies are advantageous for classification of human activities and fall detection [7] given that they are able to extract features from raw data. In that sense, we propose to use short-term memory networks (LSTM) for computing raw sensor signals from some inertial measurement units (IMU), since these type of networks can deal with subsets of time series data [8]. In addition, we are also proposing to use convolutional neural networks (CNN) mainly because they are prepared for treating image data from video cameras [9], in an easy way. Thus, we combine the two networks in order to benefit from these advantages.

The rest of the paper is organized as follows. In Section II, we reviewed related fall detection systems. Our UP-fall detection and activity recognition dataset is presented in Section III. The proposed multimodal fall detection approach is explained in Section IV. Experiments and results are shown in Section V. Finally, conclusions and future work are discussed.

This research has been funded by Universidad Panamericana through the grant “Fomento a la Investigación UP 2018”, under project code UP-CI-2018-ING-MX-04.

978-1-7281-4569-3/19/$31.00 ©2019 IEEE

II. Multimodal Fall Detection Systems

In this section, we present a review of related work. Fall detection has gained interest given aging of population and severity of the problem. Many works can be found in literature related to fall detection. For this work, we decided to review the most cited papers. We also included works that use deep learning or are addressing real-time detection.

A. Acceleration-Based Systems

Accelerometer is one of the preferred sensor for fall detection. Conventional methods and in particular threshold method, and machine learning methods are used for fall detection. In early works, authors claim that threshold is adequate for real-time processing [4] and fall dynamics identification. Nevertheless since new advanced sensors perceive more detail of human activities, threshold method is inadequate to achieve this goal [3].

In [10], data collection was done with a wireless OPAL inertial sensor attached to the waist of eleven subjects. They address the problem of near-fall or recoverable unbalances scenarios. In [11], Lin outlines wearable micro-sensing device for monitoring human falls designed as a wearable coat with embedded accelerometer based sensors. Fall detection is done with a rule-based decision tree algorithm to identify fall stages and fall detection.

Machine learning techniques have also been used for fall detection with accelerometer information. Accelerometers embedded in smartphones have gained popularity in recent years given the affordability and wide adoption of these devices. They are most commonly used and less obtrusive. Aguiar et al [12] presented an unobtrusive smartphone based approach for fall detection. Smartphone is worn in the waist or in the pocket. This works algorithm is based on a state machine that recognizes fall stages in sequential order. Features and thresholds are learned using decision trees. They presented a comparison with k-nearest neighbors (KNN) and Naive Bayes.

Kau et al. [13] also proposed an architecture for fall detection based on smartphone worn in pocket using a cascade classification of Support Vector Machine (SVM) and a state machine in order to identify changes in fall event. Pierleoni et al. [14] proposed a fall detection system based in an accelerometer placed on the waist. For classification, the authors also used a combined algorithm of threshold and training with Support Vector Machine.They are interested in identifying typical changes in acceleration during a fall and they take four phases into account occurring in falls.

B. Context-Based Systems

The most popular context-based detection systems use vision approaches, nevertheless context-aware sensors have gained interest in recent years. Feng et al. [15] presents a vision based fall detection system that analyzed four postures in a smart home environment using deep learning. They used a single camera to monitor an elderly person. They claim to prefer a deep learning classifier given that it captures interactions of many different factors on various levels through hierarchical learning.

Jokanovic et al. [16] [17] proposed a fall detection system based on deep learning using radar. Radar technologies have non-obstructive illumination, sensing in non-intrusive, they dont have privacy issues, and they are not sensitive to lighting. Features are extracted from time-frequency signature of the radar data for classification. Some Doppler signatures of different activities are similar to falls causing high false alarm rates. The authors address this issue using deep neural networks in their fall detection system taking advantage of the deep learning to learn and capture the underlying features of time frequency signature.

Another deep learning approach for fall detection based on infra red depth sensor measurements is presented in [18] . They also applied multi-layer perceptron for fall detection[19] Fan et al.[10] developed a fall detection system based on infrared thermal image temperature sensors (Grid-Eye Infrared Sensors). Their approach has two steps: a) data filtering with wavelet, Gaussian and median filters b) classification analysis with several deep learning methods including multi-layer perceptron, LSTM networks an Gated recurrent unit (GRU) networks. They compared their results with similar works [11,] that also use only Grid-Eye Infrared Sensors. In both cases, authors only consider two classes: fall and non-fall. Context-based approaches also try to identify fall dynamics with more complex machine learning methods as Deep Learning.

C. Multimodal Systems

Wang et al. [20] proposed a A fall detection system for elderly monitoring based on information from accelerometer, accelerometer, cardiotachometer and smart sensor for temperature and humidity. This system detects the motion, shock and vibration of a fall with a threshold method.

Liu te al. [21] developed a multimodal system collecting data from an IMU placed in the sternum and six-camera infrared motion capture system.They used a threshold algorithm for detecting only backwards fall. Although it claims to have a quick response to fall it is limited in fall variety detection. Motion Capture systems are usually obtrusive and difficult to implement in elderly. They also used threshold method to describe the fall dynamics which entails fall initiation, fall detection and fall completion.

Cheng et al. [22] proposed a framework for daily activity monitoring and fall detection based in surface electromyography and accelerometer signals. The authors combined sensors with the aim of better represent subtle actions and large-scale movements.

Kwolek and Kepski [23] [24] collected an interesting publicly available multimodal dataset for fall detection. Five volunteers were recorded falls and activities of daily living wearing an IMU inertial device connected via Bluetooth, and two Kinects also collected the events connected via USB. Their aim is to design a low cost and reliable fall detection system tracking a person in real-time using a detection strategy combining threshold and support vector machines (SVM). In other works, [25] they developed an algorithm applying KNN classifier for fall detection using a ceiling-mounted 3D depth camera. The authors used an accelerometer to reduce the processing overload to indicate the potential impact of the person, and to start an analysis of depth images.

Castillo et al. [26] combined video cameras, accelerometers and GPS sensors in a multimodal system for fall detection. The authors combine accelerometers for classification and video cameras to provide context for better interpretation and false-positive reduction. Decision-tree machine learning technique is used for classification.

Chen and Wang [27] proposed a sensor fusion system based on 8x pixel Grid-Eye infrared array sensor and an ultrasonic distance sensor installed on a pan-tilt orienting mechanism. This mechanism is mobile and does not have privacy issues, but it may not always work .

III. Multimodal Dataset for Fall Detection

We built a large dataset for multimodal fall detection, namely UP-Fall Detection, that includes 12 activities and three trials per activity. Data were collected over 17 subjects using a multimodal approach based on wearable sensors, ambient sensors and vision devices. Subjects performed six simple human daily activities as well as five different types of human falls. The consolidated dataset (812 GB), as well as, the feature dataset (171 GB) are publicly available in http://sites.google.com/up.edu.mx/har-up/.

During data collection, 17 subjects (9 male and 8 female) ranging from 18–24 years old, mean height of 1.66 m and mean weight of 66.8 kg, were invited to perform 11 different activities. Table I summarizes the statistics of the subjects. The activities performed are related to six simple human daily activities (walking, standing, picking up an object, sitting, jumping and laying) and five human falls (falling forward using hands, falling forward using knees, falling backwards, falling sitting in an empty chair, falling sideward and any falling finished in knees). For all the activities and falls, a mattress was located in the falling area to prevent injuries. For reliability, each activity was performed three times (trials) by each subject. Table II summarizes the activities and the duration each trial takes in the final dataset. It is important to highlight that all the subjects that participated in this dataset previously filled out an agreement with the principal investigator and the Faculty of Engineering, considering the regulations and data policies applicable.

We use five inertial measurement units as wearable sensors collecting raw data from 3-axis accelerometer, 3-axis gyroscope and ambient light value. These wearables were located in the left wrist, under the neck, at right pocket of pants, at the middle of waist (in the belt), and in the left ankle. Also, one electroencephalograph (EEG) headset was occupied to measure the raw brainwave signal from its unique EEG channel sensor located at the forehead. As context-aware sensors, we installed six infrared sensors as a grid 0.40m above the floor of the room, to measure the changes in interruption of the optical devices, where 0 means interruption and 1 no interruption. Lastly, two cameras were located at 1.82m above the floor, one for a lateral view and the other for a frontal view. Figure 1 shows the location of the wearables in the body and the layout of the context-aware sensors and cameras.

TABLE I
Statistics of the subjects

Subject No.	Age	Height (m)	Weight (kg)	Gender
1	18	1.70	99	Male
2	20	1.70	58	Male
3	19	1.57	54	Female
4	20	1.62	71	Female
5	21	1.71	69	Male
6	22	1.62	68	Male
7	24	1.74	70	Male
8	23	1.75	88	Male
9	23	1.68	70	Female
10	19	1.69	63	Male
11	20	1.65	73	Female
12	19	1.60	53	Female
13	20	1.64	55	Male
14	19	1.70	73	Female
15	21	1.57	56	Female
16	20	1.70	62	Male
17	20	1.66	54	Female

TABLE II
Activities performed by subjects

Activity No.	Description	Duration (s)
1	Falling forward using hands	10
2	Falling forward using knees	10
3	Falling backwards	10
4	Falling sitting in empty chair	10
5	Falling sideward	10
6	Walking	60
7	Standing	60
8	Sitting	60
9	Picking up an object	10
10	Jumping	30
11	Laying	60
12	In knees	–

IV. Proposed Multimodal Fall Detection Approach

We propose to use a combination of both long short-term memory (LSTM) networks and convolutional neural networks (CNN) for multimodal fall detection, as shown in Figure 2. Considering a multimodal approach with different raw signals from sensors and video recordings, our system consists on three phases: (i) a LSTM network for sensor signals, (ii) a CNN for video recordings, and (iii) a data aggregation strategy. The proposed system is described following.

A. LSTM for Sensor Signals

Falls are dynamic activities that entail a sequence of events that can be detected. An LSTM network is a type of a recurrent neural network that learns long-term dependencies over time in sequence data [8]. In that sense, LSTM networks deal with time series and can classify or predict information.

For this proposal, we use an LSTM network to deal with 42 raw sensor signals: 35 channels of wearable sensors, 1 channel of the brain wave sensor and 4 channels for infrared (context) sensors. The LSTM network architecture is depicted in Figure 2. It consists on a layer of 100 long short-term memory units, a fully connected layer of 12 units, a soft-max layer and a classification layer that recognizes the fall/activity done given an input sequence. For training purposes, we set different parameters: 250 maximum number of epochs, mini-batch size of 17, no shuffle, and an initial learning rate of 0.002.

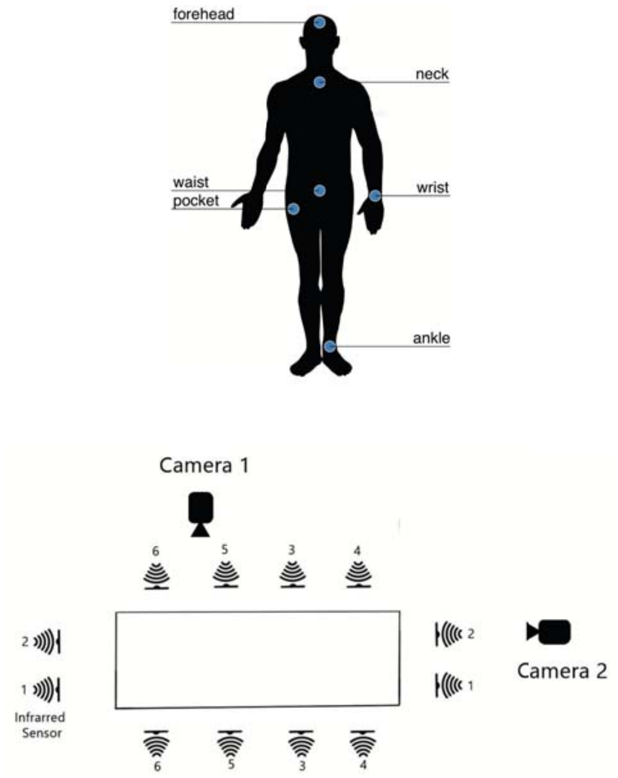


Fig. 1. Distribution of the wearable sensors, camera devices and infrared sensors.

In this proof-of-concept, the architecture and parameters were set experimentally.

B. CNN for Video Recordings

CNN are nets inspired on the nature of visual perception in living creatures mainly applied for image processing [9], [28]. Their architectures are defined using different layers, such as: convolutional, pooling and fully-connected. A convolutional layer aims to compute feature representations of the input, a pooling layer aims to reduce the resolution of feature maps, and a fully-connected layer aims to perform high-level reasoning [28]. Depending on the learning task, CNN implements an output layer aiming to compute classification or regression.

For this proposal, we use a CNN adapted for our video recordings. The proposed CNN receives as input a frame from the video recordings and estimates the fall/activity performed by the present subject. Figure 2 shows the architecture of the final CNN, and it considers the following layers: a convolutional layer with 8 filters of size 3×3 with a rectified linear unit (ReLU) and a max-pooling of size 2×2 layers; then, a convolutional layer with 16 filters of size 3×3 with a ReLU and a max-pooling of size 2×2 layers; after that, a convolutional layer with 32 filters of size 3×3 with a ReLU and a max-pooling of size 2×2 layers; and, finally, there is a fully-connected layer with output size 12 and softmax function. We trained the CNN using the stochastic gradient descent algorithm with initial learning rate of 0.001, regularization coefficient 0.004, maximum number of epochs 5, and mini-batch size of 100. Again, the architecture and parameters were selected experimentally.

C. Aggregation and Classification

To this end, we use both networks to compute an estimation of the fall/activity performed. To merge these results and output the final fall/activity detection, we obtain estimations from both networks during a temporal window size. At the end of the window, a majority voting strategy is calculated that consists on output the most frequent class within the window [29].

V. Experimentation and Results

To validate our proposal, we divide the dataset into two subsets: a training set comprises of two trials per activity and a testing set with the remaining trial per activity. For training, we consider LSTM and CNN as independent nets.

Additionally, we include a baseline classifier using random forest (RF) and applying the development cycle of a conventional fall detection system [30]. For our own dataset, we did the following steps: we windowed all raw signal data in lengths of 1 second, we extracted temporal and frequency features [31], we selected the best features, and lastly we trained the RF-model. For the images in cameras, we extracted the relative motion of pixels with optical flow [32], as features, in mean pixels from the windows. Then, we resized these features to compact information, and then we used these compact features as aggregation to the previous signal-based features to complete the feature dataset. For training purposes, we set the parameters of RF as follows: 10-trees, 2 minimum samples for splitting and bootstrap.

A. Performance of LSTM for Sensor Signals

For LSTM, we consider 14123 samples for training and 7044 samples for testing. These samples were obtained from 1-second windows of raw sensor signals with no overlapping and labelling with the most frequent fall/activity performed during that window. We ran the training process five times and we selected the best LSTM classifier using the accuracy metric over the training set. Using the testing set, the LSTM performed 96.5% of accuracy, as seen in the confusion matrix depicted in Figure 3.

From Figure 3, it can be observed that this dataset is highly unbalanced. In spite of that, LSTM perform very well on testing data. However, differentiating types of falls is a difficult task. For instance, “falling forward using knees” (2) and “falling sidewards” (5) are the less precision recognition.

B. Performance of CNN for Video Recordings

For CNN, the training data consisted on 140451 samples and the testing data on 70145 samples. Only camera 1 was used for training and testing. We ran the training process five times and we selected the best CNN classifier using the accuracy metric over the training set. After that, we validated our CNN performing 95.1% of accuracy on testing data, as shown in the confusion matrix depicted in Figure 4.

As shown in Figure 4, falls are the most difficult ones for estimating than daily activities. For instance, estimations of “falling forward using hands” (1) and “falling sideward” (5)

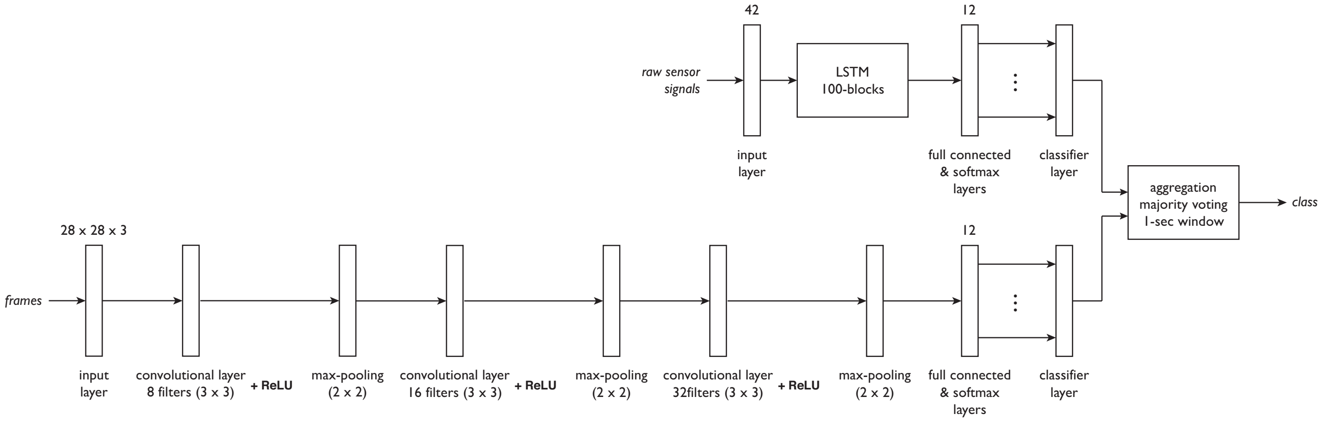
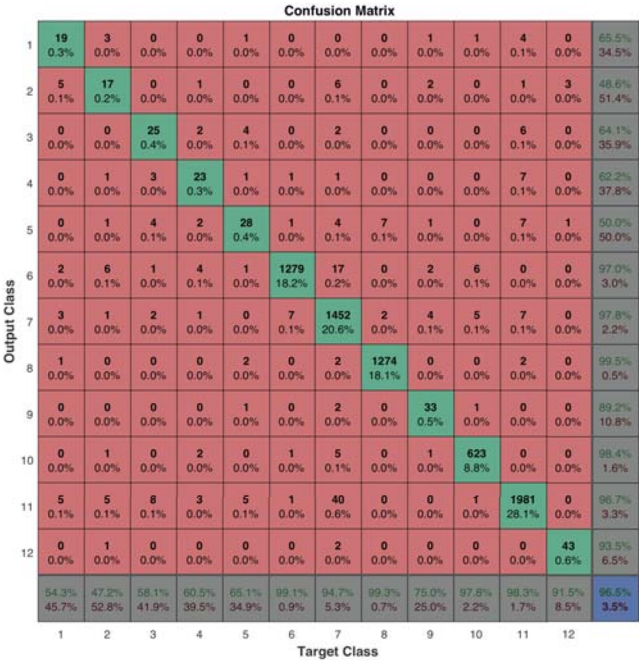
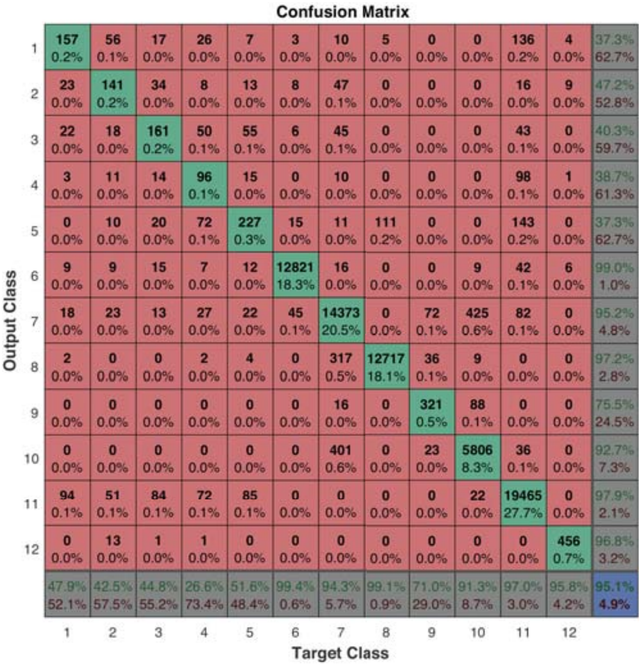

42
12
raw sensor signals
LSTM 100-blocks
input layer
full connected & softmax layers
classifier layer
aggregation majority voting 1-sec window
class
28 x 28 x 3
12
frames
input layer
convolutional layer 8 filters (3 x 3)
+ ReLU
max-pooling (2 x 2)
convolutional layer 16 filters (3 x 3)
+ ReLU
max-pooling (2 x 2)
convolutional layer 32filters (3 x 3)
+ ReLU
max-pooling (2 x 2)
full connected & softmax layers
classifier layer

Fig. 2. Block diagram of the proposed multimodal fall detection system.



	1	2	3	4	5	6	7	8	9	10	11	12	
1	19 0.3%	3 0.0%	0 0.0%	0 0.0%	1 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.0%	1 0.0%	4 0.1%	0 0.0%	65.5% 34.5%
2	5 0.1%	17 0.2%	0 0.0%	1 0.0%	0 0.0%	0 0.0%	6 0.1%	0 0.0%	2 0.0%	0 0.0%	1 0.0%	3 0.0%	48.6% 51.4%
3	0 0.0%	0 0.0%	25 0.4%	2 0.0%	4 0.1%	0 0.0%	2 0.0%	0 0.0%	0 0.0%	0 0.0%	6 0.1%	0 0.0%	64.1% 35.9%
4	0 0.0%	1 0.0%	3 0.0%	23 0.3%	1 0.0%	1 0.0%	1 0.0%	0 0.0%	0 0.0%	0 0.0%	7 0.1%	0 0.0%	62.2% 37.8%
5	0 0.0%	1 0.0%	4 0.1%	2 0.0%	28 0.4%	1 0.0%	4 0.1%	7 0.1%	1 0.0%	0 0.0%	7 0.1%	1 0.0%	50.0% 50.0%
6	2 0.0%	6 0.1%	1 0.0%	4 0.1%	1 0.0%	1279 18.2%	17 0.2%	0 0.0%	2 0.0%	6 0.1%	0 0.0%	0 0.0%	97.0% 3.0%
7	3 0.0%	1 0.0%	2 0.0%	1 0.0%	0 0.0%	7 0.1%	1452 20.6%	2 0.0%	4 0.1%	5 0.1%	7 0.1%	0 0.0%	97.8% 2.2%
8	1 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.0%	0 0.0%	2 0.0%	1274 18.1%	0 0.0%	0 0.0%	2 0.0%	0 0.0%	99.5% 0.5%
9	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.0%	0 0.0%	2 0.0%	0 0.0%	33 0.5%	1 0.0%	0 0.0%	0 0.0%	89.2% 10.8%
10	0 0.0%	1 0.0%	0 0.0%	2 0.0%	0 0.0%	1 0.0%	5 0.1%	0 0.0%	1 0.0%	623 8.8%	0 0.0%	0 0.0%	98.4% 1.6%
11	5 0.1%	5 0.1%	8 0.1%	3 0.0%	5 0.1%	1 0.0%	40 0.6%	0 0.0%	0 0.0%	1 0.0%	1981 28.1%	0 0.0%	96.7% 3.3%
12	0 0.0%	1 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	43 0.6%	93.5% 6.5%
	54.3% 45.7%	47.2% 52.8%	58.1% 41.9%	60.5% 39.5%	65.1% 34.9%	99.1% 0.9%	94.7% 5.3%	99.3% 0.7%	75.0% 25.0%	97.8% 2.2%	98.3% 1.7%	91.5% 8.5%	96.5% 3.5%

Fig. 3. Confusion matrix in testing using raw sensor signals.



	1	2	3	4	5	6	7	8	9	10	11	12	
1	157 0.2%	56 0.1%	17 0.0%	26 0.0%	7 0.0%	3 0.0%	10 0.0%	5 0.0%	0 0.0%	0 0.0%	136 0.2%	4 0.0%	37.3% 62.7%
2	23 0.0%	141 0.2%	34 0.0%	8 0.0%	13 0.0%	8 0.0%	47 0.1%	0 0.0%	0 0.0%	0 0.0%	16 0.0%	9 0.0%	47.2% 52.8%
3	22 0.0%	18 0.0%	161 0.2%	50 0.1%	55 0.1%	6 0.0%	45 0.1%	0 0.0%	0 0.0%	0 0.0%	43 0.1%	0 0.0%	40.3% 59.7%
4	3 0.0%	11 0.0%	14 0.0%	96 0.1%	15 0.0%	0 0.0%	10 0.0%	0 0.0%	0 0.0%	0 0.0%	98 0.1%	1 0.0%	38.7% 61.3%
5	0 0.0%	10 0.0%	20 0.0%	72 0.1%	227 0.3%	15 0.0%	11 0.0%	111 0.2%	0 0.0%	0 0.0%	143 0.2%	0 0.0%	37.3% 62.7%
6	9 0.0%	9 0.0%	15 0.0%	7 0.0%	12 0.0%	12821 18.3%	16 0.0%	0 0.0%	0 0.0%	9 0.0%	42 0.1%	6 0.0%	99.0% 1.0%
7	18 0.0%	23 0.0%	13 0.0%	27 0.0%	22 0.0%	45 0.1%	14373 20.5%	0 0.0%	72 0.1%	425 0.6%	82 0.1%	0 0.0%	95.2% 4.8%
8	2 0.0%	0 0.0%	0 0.0%	2 0.0%	4 0.0%	0 0.0%	317 0.5%	12717 18.1%	36 0.1%	9 0.0%	0 0.0%	0 0.0%	97.2% 2.8%
9	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	16 0.0%	0 0.0%	321 0.5%	88 0.1%	0 0.0%	0 0.0%	75.5% 24.5%
10	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	401 0.6%	0 0.0%	23 0.0%	5806 8.3%	36 0.1%	0 0.0%	92.7% 7.3%
11	94 0.1%	51 0.1%	84 0.1%	72 0.1%	85 0.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	22 0.0%	19465 27.7%	0 0.0%	97.9% 2.1%
12	0 0.0%	13 0.0%	1 0.0%	1 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	456 0.7%	96.8% 3.2%
	47.9% 52.1%	42.5% 57.5%	44.8% 55.2%	26.6% 73.4%	51.6% 48.4%	99.4% 0.6%	94.3% 5.7%	99.1% 0.9%	71.0% 29.0%	91.3% 8.7%	97.0% 3.0%	95.8% 4.2%	95.1% 4.9%

Fig. 4. Confusion matrix in testing using video recordings.

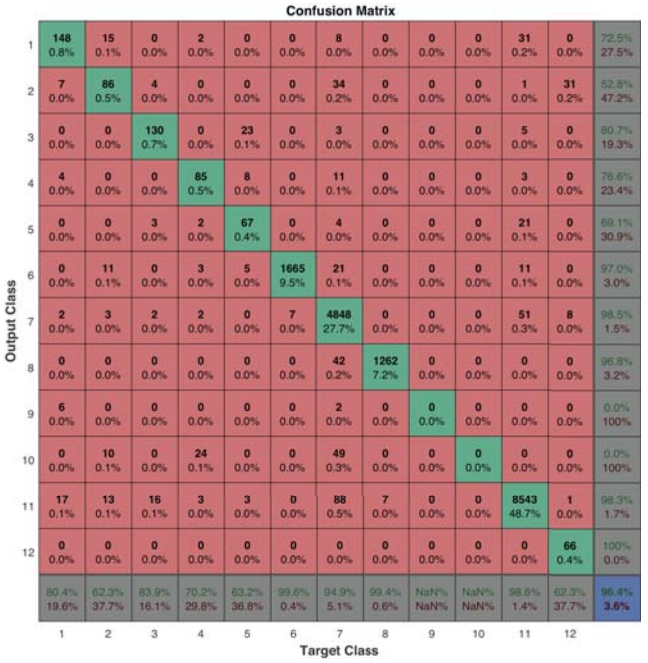
obtained the worst precision (37.3%) while "falling forward using knees" (2) obtained the largest precision (47.2%).

C. Performance of Aggregation and Final Classification

To this end, we performed the majority voting strategy in the estimations of both nets with windows of 1-second size, over a testing set of 17527 samples extracted from the third trial of all activities performed by the subjects. Figure 5 shows the confusion matrix obtained after majority voting.

As shown in the confusion matrix from Figure 5, it can be seen that the accuracy obtained was 96.4%. For a better analysis, the precision, recall and F_1-score measures were calculated and summarized in Table III. This table also reported the metrics for the RF-model baseline (i.e. feature-based training).

It can be observed that our proposal (i.e. LSTM + CNN + majority voting) improves the classification performance per activity than the single LSTM or CNN networks. For the falls, this improvement represents better precision, recall and F_1-score getting up to 50% in all the cases, in contrast to the performance of single nets. In terms of the simple activities, our proposal maintains the performance when using a single network. Notice that activities "picking up an object" (9) and "jumping" (10) were difficult to analyze since the testing set does not contain labels for them. It can be explained because activity 9 lasts less than one second and it was shadowed by the window size. In terms of activity 10, estimations failed mainly because this activity can be confused with "standing" (7) for video recordings. Finally, activity 12 computed less recall using our proposal than with the other approaches and then F_1-score was also diminished. We can observed in Figure 5 that this activity was false confused with "falling forward using knees"



Confusion Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	
1	148 0.8%	15 0.1%	0 0.0%	2 0.0%	0 0.0%	0 0.0%	8 0.0%	0 0.0%	0 0.0%	0 0.0%	31 0.2%	0 0.0%	72.5% 27.5%
2	7 0.0%	86 0.5%	4 0.0%	0 0.0%	0 0.0%	0 0.0%	34 0.2%	0 0.0%	0 0.0%	0 0.0%	1 0.0%	31 0.2%	52.8% 47.2%
3	0 0.0%	0 0.0%	130 0.7%	0 0.0%	23 0.1%	0 0.0%	3 0.0%	0 0.0%	0 0.0%	0 0.0%	5 0.0%	0 0.0%	80.7% 19.3%
4	4 0.0%	0 0.0%	0 0.0%	85 0.5%	8 0.0%	0 0.0%	11 0.1%	0 0.0%	0 0.0%	0 0.0%	3 0.0%	0 0.0%	76.6% 23.4%
5	0 0.0%	0 0.0%	3 0.0%	2 0.0%	67 0.4%	0 0.0%	4 0.0%	0 0.0%	0 0.0%	0 0.0%	21 0.1%	0 0.0%	69.1% 30.9%
6	0 0.0%	11 0.1%	0 0.0%	3 0.0%	5 0.0%	1665 9.5%	21 0.1%	0 0.0%	0 0.0%	0 0.0%	11 0.1%	0 0.0%	97.0% 3.0%
7	2 0.0%	3 0.0%	2 0.0%	2 0.0%	0 0.0%	7 0.0%	4848 27.7%	0 0.0%	0 0.0%	0 0.0%	51 0.3%	8 0.0%	98.5% 1.5%
8	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	42 0.2%	1262 7.2%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	96.8% 3.2%
9	6 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0.0% 100%
10	0 0.0%	10 0.1%	0 0.0%	24 0.1%	0 0.0%	0 0.0%	49 0.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0.0% 100%
11	17 0.1%	13 0.1%	16 0.1%	3 0.0%	3 0.0%	0 0.0%	88 0.5%	7 0.0%	0 0.0%	0 0.0%	8543 48.7%	1 0.0%	98.3% 1.7%
12	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	66 0.4%	100% 0.0%
	80.4% 19.6%	62.3% 37.7%	83.9% 16.1%	70.2% 29.8%	63.2% 36.8%	99.6% 0.4%	94.9% 5.1%	99.4% 0.6%	NaN% NaN%	NaN% NaN%	98.6% 1.4%	62.3% 37.7%	96.4% 3.6%

Output Class

Target Class

Fig. 5. Confusion matrix in testing using majority voting at 1-second size windows.

(2).

LSTM, CNN and the proposed one reached better performance than the baseline in terms of precision, recall and F$_1$-score metrics. Comparing with the baseline, our proposal computes an increment of 26.6% in precision, 18.3% in recall and 34.3% in F$_1$-score. It shows that our proposal outperforms the baseline, and it can also perform better than LSTM or CNN only networks.

Some limitations of our proposal consider that "falling forward using knees" (activity no. 2) was the most challenging fall, and "sitting" (activity no. 8) was the most difficult activity to predict, as shown in Table III. In addition, this is a preliminary work in which architectures of both LSTM and CNN were proposed experimentally, so further research in suitable architectures is required. Notice also that a majority voting approach is also desirable in this multimodal fall detection system.

To this end, it can be observed that in general our proposal performs better in accuracy, precision, recall and F$_1$-score than the RF baseline, single LSTM nets or CNN. This experiment validates our proposal. Thus, we can conclude that a multimodal approach increases the performance for fall detection.

VI. Discussion

As we reviewed in Section II, successful fall detection systems with different approaches have been presented in literature. Accelerometer based fall detection systems frequently use a threshold-based classification [20], [5], [11]. These systems depend in the location and position of sensors and wearing the sensors which is pervasive and tend to have significant number of false positives [26] given the complexity of fall dynamics. Recent fall detection systems are now using built-in sensors in smart phones [13], [12] which are easy to wear and widely adopted. Hence, the preferred position to wear a smart phone is the right trousers pocket and the orientation is still an issue. Decision Tree classifier [12] and machine learning approaches are more recently adopted as sensors evolve.

In vision-based detection systems always select machine learning classifiers [3] in order to identify fall stages and fall detection. Vision systems are not pervasive but are affected by ambient conditions and sometimes entail privacy issues. Ambient sensors and context aware systems [16], [18] try to overcome wearable obtrusiveness and vision ambient condition and privacy issues using neural networks and deep learning techniques.

Multimodal fall detection systems combining different wearable, ambient and vision sensors try to improve the performance and overcome some of the afford mentioned drawbacks. Liu et al. [21] combine IMU with a motion capture system to deal with fall dynamics, hence this approach is difficult to implement in real world systems. Multimodal approaches also tend to combine classifiers like in [23], [24] which combine threshold and machine learning techniques for classification. Our proposed multimodal detection system, collects data from five accelerometers and six infrared sensors placed in different positions. We use a LSTM network in order to deal with the dependencies over time present in fall events. It is also able to process raw signals avoiding the burden of feature extraction and selection. A CNN is also proposed to process the video recordings we retrieved from the cameras. We exploit the advantage of processing raw images with CNN, like extracting features and finding nonlinear relationships. Thus, our proposed multimodal fall detection system uses LSTM and CNN to combine their benefits for handling the nature of signals gathered in our database. We are aware that there is a difference between real falls of elderly persons and falls simulated by young healthy subjects without any impairment for safety reasons. Therefore, transfer learning approaches for prediction in elderly people or adults with impairments can be conducted.

Conclusions

In this work, we proposed to use a deep learning (LSTM and CNN) multimodal fall detection system based on wearable sensors, ambient sensors and vision devices. Particularly, this approach consisted on training LSTM and CNN models using raw inputs. To validate the approach, we built a multimodal dataset namely UP-Fall Detection that includes 12 activities performed by 17 subjects based on information collected from five inertial measurement units as wearable sensors, one electroencephalograph (EEG) headset, six infrared sensors as a grid, and two cameras.

For experimentation, we decided to use two trials of each activity performed by the subjects as training and another independent trial as testing. After training and testing, our proposal reached 96.4% in accuracy. In addition, we compared our method and the single LSTM and CNN behaviors over precision, recall and F$_1$-score metrics. After analyzing this comparison, we determined that our proposed method improves these metrics in almost every activity.

TABLE III
PERFORMANCE METRICS OF THE PROPOSED METHOD.

Activity	Precision (%)				Recall (%)				F_1-score (%)			
	Baseline (RF)	LSTM	CNN	Proposal	Baseline (RF)	LSTM	CNN	Proposal	Baseline (RF)	LSTM	CNN	Proposal
1	32.0	65.5	37.3	**72.5**	64.0	54.3	47.9	**80.4**	42.2	59.4	41.9	**76.3**
2	**53.0**	48.6	47.2	52.8	**64.6**	47.2	42.5	62.3	**58.2**	47.9	44.7	57.1
3	43.1	64.1	40.3	**80.7**	60.5	58.1	44.8	**83.9**	50.3	61.0	42.4	**82.3**
4	28.6	62.2	38.7	**76.6**	65.1	60.5	26.6	**70.2**	39.7	61.3	31.5	**73.3**
5	18.1	50.0	37.3	**69.1**	54.5	**65.1**	51.6	63.2	27.1	56.6	43.3	**66.0**
6	96.2	97.0	99.0	**97.0**	48.4	99.1	99.4	**99.6**	64.3	98.1	**99.2**	98.3
7	95.4	97.8	95.2	**98.5**	52.1	94.7	94.3	**94.9**	67.3	96.2	94.7	**96.6**
8	99.0	**99.5**	97.2	96.8	92.5	99.3	99.1	**99.4**	95.6	**99.4**	98.1	98.1
9	42.1	89.2	75.5	–	85.7	75.0	71.0	–	56.4	81.5	73.2	–
10	97.3	98.4	92.7	–	82.9	97.8	91.3	–	89.5	98.1	92.0	–
11	**99.3**	96.7	97.9	98.3	58.2	98.3	97.0	**98.6**	73.3	97.5	97.5	**98.4**
12	93.7	93.5	96.8	**100**	**98.9**	91.5	95.8	62.3	96.2	92.5	**96.3**	76.7
mean	66.5	80.2	71.3	**84.2**	68.9	78.4	71.8	**81.5**	61.3	79.1	71.2	**82.3**

For future work, we are considering to test other aggregation and fusion methods. Also, we will test the performance of our method on a leave-one-subject-out validation to determine the robustness of the proposal on new subjects.

DATA DISCLOSURE

UP-Fall Detection Dataset is publicly available at: http://sites.google.com/up.edu.mx/har-up/. The database website will be provisionally limited from December 3, 2018 to July 19, 2019 due to an open competition that uses this dataset. If interested on using it, we encourage users to contact the correspondence authors for data accessibility during this period.

REFERENCES

[1] World Health Organization, “Falls,” https://www.who.int/news-room/fact-sheets/detail/falls, 2018, [Online; accessed 13-Dec-2018].

[2] R. Igual, C. Medrano, and I. Plaza, “Challenges, issues and trends in fall detection systems,” *Biomedical engineering online*, vol. 12, no. 1, p. 66, 2013.

[3] T. Xu, Y. Zhou, and J. Zhu, “New advances and challenges of fall detection systems: A survey,” *Applied Sciences*, vol. 8, no. 3, p. 418, 2018.

[4] D. M. Karantonis, M. R. Narayanan, M. Mathie, N. H. Lovell, and B. G. Celler, “Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring,” *IEEE transactions on information technology in biomedicine*, vol. 10, no. 1, pp. 156–167, 2006.

[5] L. Liu, M. Popescu, M. Skubic, M. Rantz, T. Yardibi, and P. Cuddihy, “Automatic fall detection based on doppler radar motion signature,” in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2011 5th International Conference on*. Citeseer, 2011, pp. 222–225.

[6] J. T. Perry, S. Kellog, S. M. Vaidya, J.-H. Youn, H. Ali, and H. Sharif, “Survey and evaluation of real-time fall detection approaches,” in *High-Capacity Optical Networks and Enabling Technologies (HONET), 2009 6th International Symposium on*. IEEE, 2009, pp. 158–164.

[7] D. Ravi, C. Wong, B. Lo, and G.-Z. Yang, “A deep learning approach to on-node sensor data analytics for mobile or wearable devices,” *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 56–64, 2017.

[8] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, “Skeleton-based human action recognition with global context-aware attention lstm networks,” *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1586–1599, 2017.

[9] K. Nogueira, O. Penatti, and J. dos Santos, “Towards better exploiting convolutional neural networks for remote sensing scene classification,” *Pattern Recognition*, vol. 61, pp. 539 – 556, 2017.

[10] J. K. Lee, S. N. Robinovitch, and E. J. Park, “Inertial sensing-based pre-impact detection of falls involving near-fall scenarios,” *IEEE transactions on neural systems and rehabilitation engineering*, vol. 23, no. 2, pp. 258–266, 2015.

[11] C.-S. Lin, H. C. Hsu, Y.-L. Lay, C.-C. Chiu, and C.-S. Chao, “Wearable device for real-time monitoring of human falls,” *Measurement*, vol. 40, no. 9-10, pp. 831–840, 2007.

[12] B. Aguiar, T. Rocha, J. Silva, and I. Sousa, “Accelerometer-based fall detection for smartphones,” in *Medical Measurements and Applications (MeMeA), 2014 IEEE International Symposium on*. IEEE, 2014, pp. 1–6.

[13] L.-J. Kau and C.-S. Chen, “A smart phone-based pocket fall accident detection, positioning, and rescue system,” *IEEE journal of biomedical and health informatics*, vol. 19, no. 1, pp. 44–56, 2015.

[14] P. Pierleoni, A. Belli, L. Palma, M. Pellegrini, L. Pernini, and S. Valenti, “A high reliability wearable device for elderly fall detection,” *IEEE Sensors Journal*, vol. 15, no. 8, pp. 4544–4553, 2015.

[15] P. Feng, M. Yu, S. M. Naqvi, and J. A. Chambers, “Deep learning for posture analysis in fall detection,” in *Digital Signal Processing (DSP), 2014 19th International Conference on*. IEEE, 2014, pp. 12–17.

[16] B. Jokanovic, M. Amin, and F. Ahmad, “Radar fall motion detection using deep learning,” in *Radar Conference (RadarConf), 2016 IEEE*. IEEE, 2016, pp. 1–6.

[17] B. Jokanovic, M. G. Amin, and F. Ahmad, “Effect of data representations on deep learning in fall detection,” in *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2016 IEEE*. IEEE, 2016, pp. 1–5.

[18] S. Jankowski, Z. Szymański, U. Dziomin, P. Mazurek, and J. Wagner, “Deep learning classifier for fall detection based on ir distance sensor data,” in *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2015 IEEE 8th International Conference on*, vol. 2. IEEE, 2015, pp. 723–727.

[19] S. Jankowski, Z. Szymański, P. Mazurek, and J. Wagner, “Neural network classifier for fall detection improved by gram-schmidt variable selection,” in *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2015 IEEE 8th International Conference on*, vol. 2. IEEE, 2015, pp. 728–732.

[20] J. Wang, Z. Zhang, B. Li, S. Lee, and R. S. Sherratt, “An enhanced fall detection system for elderly person monitoring using consumer home networks,” *IEEE transactions on consumer electronics*, vol. 60, no. 1, pp. 23–29, 2014.

[21] J. Liu and T. E. Lockhart, “Development and evaluation of a prior-to-impact fall event detection algorithm,” *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 7, pp. 2135–2140, 2014.

[22] J. Cheng, X. Chen, and M. Shen, “A framework for daily activity monitoring and fall detection based on surface electromyography and accelerometer signals,” *IEEE journal of biomedical and health informatics*, vol. 17, no. 1, pp. 38–45, 2013.

[23] B. Kwolek 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.

[24] ——, “Improving fall detection by the use of depth sensor and accelerometer,” *Neurocomputing*, vol. 168, pp. 637–645, 2015.

[25] M. Kepski and B. Kwolek, “Fall detection using ceiling-mounted 3d depth camera,” in *2014 International conference on computer vision theory and applications (VISAPP)*. IEEE, 2014, pp. 640–647.

[26] J. C. Castillo, D. Carneiro, J. Serrano-Cuerda, P. Novais, A. Fernández-Caballero, and J. Neves, “A multi-modal approach for activity classification and fall detection,” *International Journal of Systems Science*, vol. 45, no. 4, pp. 810–824, 2014.

[27] Z. Chen and Y. Wang, “Infrared–ultrasonic sensor fusion for support vector machine–based fall detection,” *Journal of Intelligent Material Systems and Structures*, vol. 29, no. 9, pp. 2027–2039, 2018.

[28] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, “Recent advances in convolutional neural networks,” *Pattern Recognition*, vol. 2017, pp. 1–24, 2017.

[29] A. Bayat, M. Pomplun, and D. A. Tran, “A study on human activity recognition using accelerometer data from smartphones,” *Procedia Computer Science*, vol. 34, pp. 450–457, 2014.

[30] A. Bulling, U. Blanke, and B. Schiele, “A tutorial on human activity recognition using body-worn inertial sensors,” *ACM Computing Surveys (CSUR)*, vol. 46, no. 3, pp. 1–33, 2014.

[31] H. Ponce, L. Miralles-Pechuán, and L. Martínez-Villasenor, “A flexible approach for human activity recognition using artificial hydrocarbon networks,” *Sensors*, vol. 16, no. 11, p. 1715, 2016.

[32] B. K. P. Horn and B. G. Schunck, “Determining optical flow,” *Artificial Intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.