IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 4, 2021, accepted February 13, 2021, date of publication February 23, 2021, date of current version March 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3061626

Fall Detection and Activity Recognition Using Human Skeleton Features

HEILYM RAMIREZ[1], SERGIO A. VELASTIN[2,3], (Senior Member, IEEE), IGNACIO MEZA[1], ERNESTO FABREGAS[4], DIMITRIOS MAKRIS[5], AND GONZALO FARIAS[1]

[1]Escuela de Ingeniería Eléctrica, Pontificia Universidad Católica de Valparaíso, Valparaíso 2362804, Chile
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K.
[3]Department of Computer Science and Engineering, Universidad Carlos III de Madrid, 28911 Leganés, Spain
[4]Departamento de Informática y Automática, Universidad Nacional de Educación a Distancia, 28040 Madrid, Spain
[5]Faculty of Science, Engineering and Computing, Kingston University, London SW15 3DW, U.K.

Corresponding author: Gonzalo Farias (gonzalo.farias@pucv.cl)

This work was supported in part by the Chilean Ministry of Education under Project FONDECYT 1191188.

ABSTRACT Human activity recognition has attracted the attention of researchers around the world. This is an interesting problem that can be addressed in different ways. Many approaches have been presented during the last years. These applications present solutions to recognize different kinds of activities such as if the person is walking, running, jumping, jogging, or falling, among others. Amongst all these activities, fall detection has special importance because it is a common dangerous event for people of all ages with a more negative impact on the elderly population. Usually, these applications use sensors to detect sudden changes in the movement of the person. These kinds of sensors can be embedded in smartphones, necklaces, or smart wristbands to make them ''wearable'' devices. The main inconvenience is that these devices have to be placed on the subjects' bodies. This might be uncomfortable and is not always feasible because this type of sensor must be monitored constantly, and can not be used in open spaces with unknown people. In this way, fall detection from video camera images presents some advantages over the wearable sensor-based approaches. This paper presents a vision-based approach to fall detection and activity recognition. The main contribution of the proposed method is to detect falls only by using images from a standard video-camera without the need to use environmental sensors. It carries out the detection using human skeleton estimation for features extraction. The use of human skeleton detection opens the possibility for detecting not only falls but also different kind of activities for several subjects in the same scene. So this approach can be used in real environments, where a large number of people may be present at the same time. The method is evaluated with the UP-FALL public dataset and surpasses the performance of other fall detection and activities recognition systems that use that dataset.

INDEX TERMS Fall detection, deep learning, human skeleton.

I. INTRODUCTION

Human activity recognition has attracted the attention of researchers in recent years. A significant number of works has been reported in the literature related to these topics [1], [2]. Many such studies show that this is an interesting field to apply machine learning approaches [3]. Some of the activities that can be detected using different techniques and sensors include walking [4], running [5], standing [6], sitting [7], jumping [8] and falling [9].

In this context, the detection of falls has special relevance because of the number of daily life applications it includes. It is well known that the risk of falling increases with age. Falls are considered one of the main causes of serious (or even fatal) injuries with a higher incidence in older people [10]. They often cause painful injuries that are costly because they are difficult to treat and heal. In 2015, fatal falls by the elderly had an approximate cost of 754 million dollars in the United States alone [11].

Some approaches use sensors for fall detection, such as accelerometers, barometers, inertial sensors, and gyroscopes. These kinds of sensors can be embedded in smartphones, necklaces, or smart wristbands to make them ''wearable''

The associate editor coordinating the review of this manuscript and approving it for publication was Qingli Li.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

devices. Other sensors are placed on the waist or on the chest [12] of subjects, to simplify the detection of a fall. However, they have the disadvantage that the devices have to be placed on the subjects' bodies. It can be uncomfortable and is not always feasible as this type of sensor must be worn constantly [13], [14].

Other applications use depth images obtained by Kinect-like cameras and depth sensors [15], [16], which provide the distance to the detected object using an infrared sensor (IR) [17]. Some researchers have used the fusion of images from video-cameras with data from environmental sensors (infrared, ultrasonic, etc) such as [18]–[21]. All these kinds of systems allow the segmentation of the human body on the image and they provide contactless and non-intrusive monitoring, which is an advantage for practical deployment and users' acceptance and compliance, compared with other sensor technologies, like wearable devices [22].

Using a computer vision approach, falls can be recognized by processing the images and detecting the movement of the human body by means of different machine learning algorithms such as K-nearest neighbor [23], Recurrent Neural Networks (RNNs) [24], Convolutional Neural Networks (CNNs) [25] and Support Vector Machines (SVMs) [26]. The main advantage of these solutions is that they are contactless, non-intrusive, and many subjects could be monitored in the scene at the same time.

This paper presents a vision-based approach for fall detection and activity recognition. The main contribution of the proposed method is to detect falls only by using images from a standard video-camera without the need to use depth (such as Kinect cameras) or environmental sensors. A novel aspect it that it carries out the detection using human skeleton estimation for features extraction. The use of the human skeleton detection opens the possibility for detecting not only falls but also different kinds of activities for several subjects in the same scene. So this approach can be useful for real environments, where a large number of people may be acting at the same time.

The proposed approach has been validated with the public multi-modal UP-FALL dataset presented in [27]. Four different machine learning algorithms have been tested. The results outperform by a significant margin those obtained by other methods tested on the same dataset [27].

The rest of the paper is structured as follows. Section II presents some fall detection approaches that can be found in the literature and describes briefly the UP-FALL dataset, which is the start point of this work. Section III details the proposed fall detection and activity recognition approach. Section IV shows the experimental results and a comparison with previous results for the UP-FALL dataset. Finally, Section V summarizes the main conclusions and future work.

II. FALL DETECTION DATASETS AND RELATED WORK

In this section, several approaches for fall detection described in the literature are discussed. After that, the UP-Fall dataset, which has been chosen to test the proposed approach and compare it to other reported methods, is also described.

A. FALL DETECTION APPROACHES

A considerable number of applications and datasets for fall detection have been reported in the literature. For example, SDUFALL [28] presents a dataset obtained with a Kinect camera. It includes five daily living activities and falls performed by ten young women and men. Actions are simulated and they include some changes such as carrying/not carrying an object, lights on/off, changes of position and direction relative to the camera. The authors report an accuracy of 79.91%. Other works, which use the same dataset, report similar or better accuracy. For example, [35] proposes a fall detection approach aiming to build a support system for elderly people living alone in their homes using depth videos, showing a 100% accuracy for falls and 80% for the absence of falls. In [36] they reach a fall detection accuracy of up to 92.98% based on a depth camera and using human shape and motion. [37] presents a human skeleton based fall detection method for industrial settings using LSTM and show that the proposed method is more efficient in detecting falls compared to using raw skeletal data.

The work presented in [29] shows two datasets recorded with two Kinect cameras simultaneously from two different points of view. In the first dataset (EDF), 10 subjects performed 2 falls for each of eight directions in each point of view. They also recorded five more different actions that could be like falling: picking up something, sitting on the floor, laying, tying shoelaces, do plank exercise. The second dataset (OCCU) is focused on collecting occluded falls. Five subjects performed 60 occluded falls and similar different actions as in the first dataset.

URFALL [30] is a dataset that was generated by collecting data from an IMU inertial device connected via Bluetooth and 2 Kinect cameras. Five volunteers were recorded doing 70 fall sequences. Some of these are fall-like activities in typical rooms. There were two kinds of falls: falling from a standing position and falling from sitting on a chair. Each record contains sequences of depth and RGB images for two cameras and raw accelerometer data. The authors used a threshold-based fall detection method and report an accuracy of 94.99%. Other work using this dataset include [35] achieving 100% accuracy for falling activities and 82.50% for non-falling, while [38] reports an accuracy of 97.33% detecting falls based on the integration of two data sources. Object detection (Yolo) and a human posture detection model (OpenPose) are used for pre-processing to obtain key points and position information of a human body. Meanwhile, [39] presents a method based on prominence maps showing an accuracy of 99.67% accuracy in the detection of URFALL falls. Prominence detection uses a two-stream convolutional neural network that extracts global and local features to generate the prominence maps.

The work presented in [31] provides a dataset containing RGB-D and body-worn motion-capture data for a person

performing daily life activities in a scene with occlusions. This dataset aims to provide a novel benchmark for the evaluation of different human body pose estimation systems in challenging situations. It is the first RGB-D dataset that provides ground truth data for different body parts of a person moving in a scene with occlusions. The challenge presented by this benchmark is to test the ability of a tracking system to handle severe occlusions and resume tracking when the person is again fully visible.

The research presented in [32] describes a dataset composed of ADL (activity daily living) and fall actions simulated by 11 volunteers. The people involved in the test are aged between 22 and 39, with different heights ($1.62 - 1.97\ m$) and build. The actions performed by a single person are separated into two main groups: ADL and Fall. Each activity is repeated three times by each subject involved.

PKU-MMD [33] is a recent large-scale benchmark for continuous multi-modality 3D human action understanding and covers a wide range of complex human activities with well-annotated information. It contains two phases for action detection tasks with increasing difficulty. Phase 1 is a large-margin action detection task. Phase 2 is a small-margin action detection task. The dataset also provides multi-modality data sources, including RGB, depth, Infrared Radiation and Skeleton. This large-scale dataset can benefit future research on action detection for the community. The authors test four models: JCRRNN, SVM, BLSTM and STA-LSTM, achieving F_1-Scores of 52.6%, 13.1%, 33.3% and 31.6% respectively.

PoseTrack [40] is a large-scale benchmark for human pose estimation and articulated tracking in video. It provides a publicly available training and validation set and an evaluation server for benchmarking on a held-out test set. The benchmark is a basis for the challenge competitions at ICCV'17 and ECCV'18 PoseTrack workshops with a typical accuracy of around 66% for a Faster R-CNN model.

NTU RGB-D [34] provides two datasets that contain RGB videos, depth map sequences, 3D skeletal data, and infrared (IR) videos for each sample. Each of them is captured by three Kinect V2 cameras concurrently. The actions in these datasets are divided into three major categories: daily actions, mutual actions, and medical conditions. The authors present five accuracy results: Spatio-Temporal LSTM (STLSTM) obtaining 57.9%, Part-Aware LSTM 26.3%, Soft RNN 44.9%, Multi-Task CNN with RotClips yields 61.8% and FSNet 62.4%. In [41] a 3D skeleton-based fall detection system for deep learning technique is described, showing results of the high precision and robustness of the NTU RGB-D reference data set. The proposed system has been implemented on the NVIDIA Jetson Tx2 platform with real-time processing. On the other hand, the method presented in [42] achieves an accuracy of 91.7% using OPENPOSE to obtain articulated bodies. Transfer learning is then used to train the dataset and obtain a new model that predicts the fall.

UP-FALL [27] is a large dataset mainly for fall detection, that includes 11 activities and 3 trials per activity. Subjects performed six simple human daily activities and five different human falls. This data was collected with 17 healthy young adults without physical impairment using a multimodal approach, i.e. wearable sensors, ambient sensors and vision devices. Table 6 shows results reported in [27] UP-FALL.

Using UP-FALL, [43] compares three machine learning approaches: Decision Tree, XGBoost, and Random Forest, obtaining an accuracy of 98% and an F1 of 82.47%, reaching top place in the UP-Multimodal Fall Detection Challenge [44] held in 2019.

Table 1 summarizes the main characteristics of these datasets.

B. BASELINE RESULTS: UP-FALL DATASET

The work presented here uses the UP-Fall dataset (https://sites.google.com/up.edu.mx/har-up/) for direct comparisons against other techniques. The dataset is multimodal with data captured using five Mbientlab MetaSensor wearable sensors (IMU), one electroencephalograph (EEG) NeuroSky MindWave headset and six infrared sensors (IR). It also includes data obtained from two Microsoft LifeCam Cinema cameras (CAM) placed at 1.82 m above the floor, one for a lateral view and the other for a frontal view. UP-FALL includes six simple human daily activities (walking, standing, lifting an object, sitting, jumping and lying down) and five human falls (falling forward with hands, falling forward with knees, falling backward, landing sitting on an empty chair and fall sideways), as indicated in Table 2. The instances of non-fall activities outnumber falls (approximately in a ratio of 3.5:1) so as to make it more representative of the sporadic nature of fall events.

There are two separate challenges associated with the dataset. First, fall detection as a binary classification problem to distinguish between a fall (any of classes 1 to 5) and a non-fall (any of the remaining classes). Secondly, what in the original paper [27] and those that also use this dataset generally called 'activity recognition': a multi-class classification problem of detecting each of the 12 classes in Table 2 separately. To avoid any confusion, we have kept the same terminology, although 'activity recognition' is normally used for a much wider range of actions.

The original work on [27] evaluates four different well-known machine learning (ML) methods for the fall detection and activity recognition problems: Random Forest (RF), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP) and k-Nearest Neighbors (kNN). They used different combinations of sensing modalities:

1) Infrared sensors (IR).
2) Wearable IMUs (IMU).
3) All wearable IMUs plus the EEG headset (IMU+EEG).
4) All infrared sensors, all wearable IMUs and the EEG headset (IR+IMU+EEG).
5) Video cameras (CAM).
6) All infrared sensors and video cameras (IR+CAM).

TABLE 1. Vision-based datasets for fall detection.

dataset	Camera Type	Subjects	Fall Types	Other Activities	Trials	Variants	ML Method	Performance
SDUFall [28]	1 Kinect	10	Fall to the floor	Sitting, walking, squatting, lying, bending	6 actions 10 times	Carrying, not carrying, light on, light off, position and direction changes	Bag of words model built upon curvature scale space features	Accuracy: 79.91%, sensitivity 81.91%, specificity 76.62%
EDF-OCCU [29]	2 Kinect depth	5 to 10	Falls in 8 different directions	Laying, picking up, sitting (floor), tying shoelaces, plank exercise	5 actions/ 20 times	Different directions and occluded falls		
URFall [30]	2 Microsoft Kinect	5	From standing, from sitting on a chair	Lying, walking, sitting down, crouching down	70 sequences		Support vector machine, k-nearest neighbor	Accuracy: 94.99%, precision 89.57%, sensitivity 100%, specificity 91.25%
LARSEN [31]	8 Infra-red cameras			Moving, sitting, walking, joking...	12 sequences	With occlusions	Hidden markov model	
TST v2 [32]	1 Microsoft Kinect v2	11	From front, backward, to the side	Sitting, grasp an object from the floor, walking lying	264 sequences		Depth frame	
PKU-MMD [33]	3 Microsoft Kinect v2	66		Drinking, waving hand, putting on the glassed, hugging, shaking...	6 sequences	10 interaction actions	Regression RNN Regression SVM Bidirectional LSTM S-T attention LSTM	F_1-Score: 52.6% 13.1% 33.3% 31.6%
NTU RGB-D [34]	3 Microsoft Kinect v2	Multiple actors	Fall to the floor	120 actions		Captured from 155 different camera viewpoints	S-T LSTM Part-Aware LSTM RNN CNN FSNet	Accuracy: 57.9% 26.3% 44.9% 61.8% 62.4%
UP-FALL [27]	2 cameras	17	Forward using hands, forward using knees, backwards, sideward, sitting	Walking, standing, sitting picking up an object jumping Laying	3 repetitions		Random forest SVM NN k-nearest neighbor	Accuracy: 31.96%, precision 13.04%, recall 13.73%, specificity 72.05%, F_1-Score 12.68%

7) All wearable IMUs, EEG headset and video cameras (IMU+EEG+CAM).

One of their findings is that using video-only (CAM) data gave poor results compared to the other modalities. These were improved significantly when using a convolutional neural network (CNN). kNN produced the best activity recognition results for the video sensors (CAM) with accuracy of 34.03% and F_1-Score of 15.19%.

III. PROPOSED METHOD

This work focuses on improving the performance of fall detection and activity recognition using only the video data, as in practical applications such as assisted living and public space monitoring, the use of wearables and other sensor modalities may not be feasible. The main hypothesis is that the results can be improved significantly by using articulated bodies (skeletons) extracted from the video, even when using the same ML methods used by the baseline. The workflow developed for this study is shown in Fig. 1.

A. FEATURE EXTRACTION PROCESS

1) HUMAN SKELETON DETECTION

Human pose detection is done by using AlphaPose [45], an open source method for multi-person pose estimator

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3061626, IEEE Access

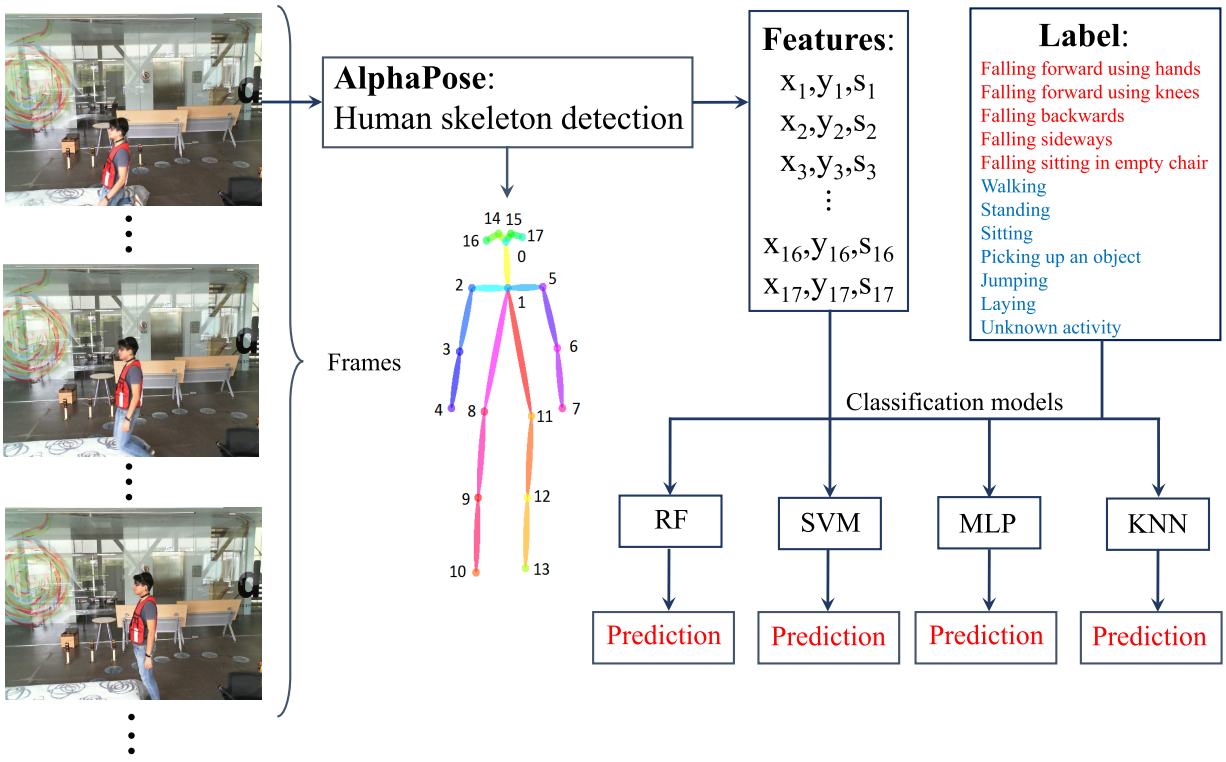

FIGURE 1. Workflow for activity recognition.

TABLE 2. UP-FALL Activities and their corresponding IDs.

Activity ID	Description
1	Falling forward using hands
2	Falling forward using knees
3	Falling backwards
4	Falling sideways
5	Falling sitting in empty chair
6	Walking
7	Standing
8	Sitting
9	Picking up an object
10	Jumping
11	Laying
20	Unknown activity

available from https://www.mvig.org/research/alphapose.html. It uses RGB images as input, then performs pose detection with a pre-trained model (COCO dataset), outputting for each processed image. For each detected person, there is an overall detection score plus a set of 17 keypoints or joints, with an individual joint score s and coordinates (x,y), which when joined form a skeleton. With these 51 (17×3) attributes, the characteristics are obtained to train a classifier to detect falls. It should be noted that the method presented here is not dependent on AlphaPose and any pose estimation method that estimates joint positions can be used. In this manner, a sequence of RGB images is converted into a sequence of (skeleton) joints and scores which forms the vector features used to learn to distinguish the different actions. Fig. 1 summarizes the proposed approach. Typical examples of skeleton detection are shown in Fig. 2. To allow a direct comparison with other works, the data from CAM1 (side view) of the UP-FALL dataset has been chosen for the experiments.

2) PRE-TRAINING FILTERING

Before any training is undertaken, the skeleton sequences are pre-processed to first remove empty frames. Secondly, in some images other people appear, in addition to the volunteer carrying out the action and to whom the action ground-truth action label applies. Such people can be passers-by or other volunteers not involved in the action. A typical example is shown in Fig. 3. It was found that the main actor can be identified by choosing the skeleton with the highest overall score and in that way eliminating all the others from the training process.

B. DATA-DRIVEN MODELS

The results of the proposed approach have been validated using the same experimental methodology described in [27]. Experiments were performed using 70% of the UP-FALL dataset for training and the remaining 30% for testing. Similarly, ten rounds of cross-validation were carried out with four classification methods: Random forest (RF), Support Vector Machine (SVM), Multilayer Perceptron (MLP) and K-Nearest Neighbor (KNN). Table 3 shows the parameters settings for each ML model, which are the same used in [27].

C. EVALUATION METRICS

To allow a direct comparison, this work uses the same performance metrics used in [27], i.e. accuracy, precision, sensitivity, specificity and F1-score. Where:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3061626, IEEE Access

FIGURE 2. Typical estimated skeletons.

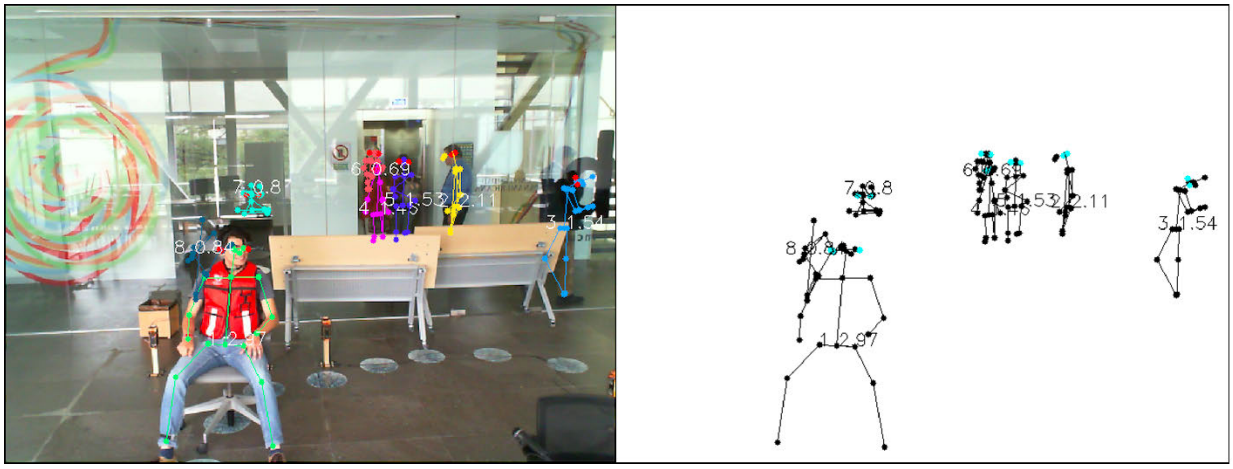

FIGURE 3. An example of a multi-person image (the person carrying out the action is in the foreground).

- True positives (TP): "Fall" detected as "Fall".
- False positives (FP): "Not Fall" detected as "Fall".
- True negatives (TN): "Not Fall" detected as "Not Fall".
- False negatives (FN): "Fall" detected as "Not Fall".

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

$$specificity = \frac{TN}{TN + FP} \tag{4}$$

Finally, F_1-Score is calculated as shown in Eq. 5 and is used to evaluate with a single value the combination of both precision and recall.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{5}$$

IV. EXPERIMENTAL RESULTS

A. FALL DETECTION

Fall detection is performed using binary classifiers. The inputs are the images from the CAM1 set, so as to compare it directly with other reported work on UP-FALL. There are 220,660 images of which 49,544 correspond to falls and 171,116 to non-falls. 2D skeleton coordinates for each image are then obtained with AlphaPose, although any pose estimation outputting 2D or 3D articulated body coordinates could be used. The data samples are separated assigning 70% for training (154,462 samples) and 30% (66,198 samples) for testing, the same split used in comparable work. The original twelve class labels are converted so that the five types of falls are coded as "Fall" and the rest as "Not Fall". Four classifiers are trained separately (RF, SVM, MLP and KNN). Performance evaluation is carried out using ten rounds (k-fold = 10) of cross-validation using random 70:30 partitions of the whole dataset. Each pose frame is processed independently to detect whether it represents a fall. Table 4 shows

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3061626, IEEE Access

TABLE 3. Parameter settings for the different ML-models.

ML-Model	Parameters
Random forest (RF)	estimators = 10 min. samples split = 2 min. samples leaf = 1 bootstrap = true
Support Vector Machine (SVM)	c = 1.0 min. samples split = 2 kernel = radial basis function kernel coefficient = 1/features shrinking = true tolerance = 0.001
Multilayer Perceptron (MLP)	hidden layer size = 100 activation function = ReLU solver = stochasticgradient penalty parameter = 0.0001 batch size = min (200, samples) initial learning rate = 0.001 shuffle = true tolerance = 0.0001 exponential decay (1st moment) = 0.9 exponential decay (2nd moment) = 0.999 regularization coefficient = $1e^{-8}$ max. epochs = 10
K Nearest Neighbor (KNN)	neighbors = 5 leaf size = 30 metric = Euclidean

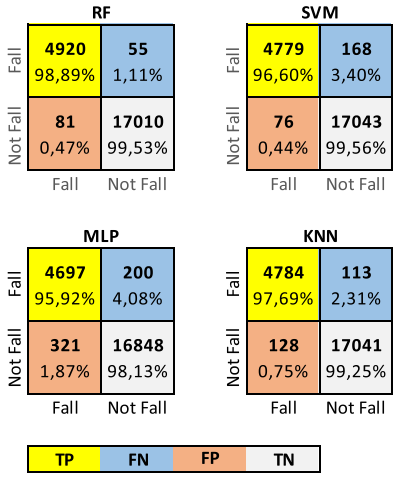

FIGURE 4. Confusion matrices for fall detection for RF, SVM, MLP and KNN respectively.

the results obtained by the four classifiers. The best results are obtained with RF reaching a mean accuracy of 99.34% and an F_1-Score of 98.52%. Considering the four models (see Table 4), the average accuracy is 98.59%, average precision 96.94%, average recall 96.80%, average specificity 99.11% and average F_1-Score is 96.87%. This compares very favorably with the results reported in [27] and even in more recent works that use the same dataset [46] where they get a F_1-Score of 96.6% with their best model.

Figure 4 shows the best confusion matrix for each classifier. It is observed in the confusion matrix of the RF model that of all the fall data only 55 cases (1.1%) are not recognized as falls.

Therefore, it can be concluded that the original hypothesis is demonstrated, namely that it is possible to detect falls with

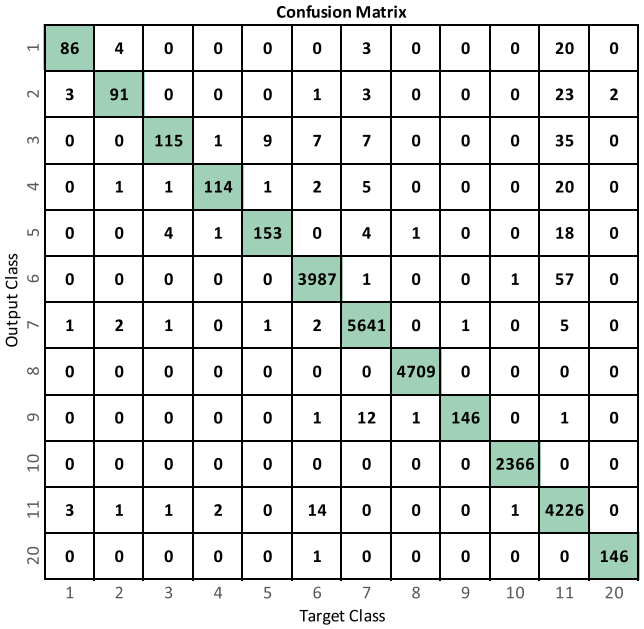

FIGURE 5. Confusion matrix for RF.

the four classifiers proposed in [27] when using only a single modality (vision from a camera) through the use of human skeleton features, obtained via deep neural models, which considerably improves the performance of these models.

B. ACTIVITY RECOGNITION

As pointed out earlier, UP-FALL defines as Activity Recognition the task of classifying a finer grain of 12 separate cases (or activities, including an 'unknown' class) as shown in Table 2. As before, the input is the set of CAM1 images that are converted into pose frames through AlphaPose. As was the case for fall detection, the dataset is randomly partitioned 10 times, into 70% for training and 30% for testing, following the scheme used in [27]. For each fold and each classifier, metrics are computed and their means and standard deviations are calculated.

Table 5 shows the results obtained by the four classifiers. SVM has the best performance for accuracy and sensitivity while for all the other metrics, the best results are obtained by the RF classifier, which also shows more uniformity in its results.

Fig. 5 shows the confusion matrix for the best classifier. It can be seen that RF activities 8 (sitting) and 10 (jumping) are detected with an accuracy of 100% and the rest of the activities are recognised with accuracies between 90.67% and 99.77%.

It is possible to see from the confusion matrix that the first 5 activities, defined as falls in UP-FALL (Table 2), are confused with activity 11 (laying). This is because in the present work the analysis is carried out for each frame individually and not for sequences of frames. This means that in a fall, when a person is about to lie on the floor it is very difficult to distinguish between such actions from a

TABLE 4. Performance (mean ± standard deviation) obtained for each model of the proposed fall detection system. Best results are shown in bold.

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F_1-Score (%)
RF	**99.34 ± 0.03**	**98.23 ± 0.17**	**98.82 ± 0.10**	**99.48 ± 0.05**	**98.52 ± 0.08**
SVM	98.81 ± 0.07	98.15 ± 0.19	96.50 ± 0.27	99.47 ± 0.05	97.32 ± 0.17
MLP	97.39 ± 0.10	93.87 ± 0.85	94.57 ± 1.15	98.21 ± 0.29	94.21 ± 0.27
KNN	98.84 ± 0.06	97.53 ± 0.15	97.30 ± 0.24	99.29 ± 0.04	97.41 ± 0.16
Average	98.59 ± 0.06	96.94 ± 0.34	96.80 ± 0.44	99.11 ± 0.11	96.87 ± 0.17

TABLE 5. Performance (mean ± standard deviation) obtained for each model of the activity recognition system.

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F_1-Score (%)
RF	99.45 ± 1.02	**96.60 ± 0.48**	**88.99 ± 0.56**	99.70 ± 0.50	**92.34 ± 0.39**
SVM	**99.65 ± 0.01**	93.85 ± 0.65	87.29 ± 0.83	**99.79 ± 0.01**	90.20 ± 0.59
MLP	98.93 ± 0.17	85.39 ± 1.69	71.44 ± 2.30	99.34 ± 0.11	75.95 ± 1.84
KNN	99.60 ± 0.01	91.65 ± 0.55	84.17 ± 0.81	99.76 ± 0.01	87.35 ± 0.63
Average	99.41 ± 0.30	91.87 ± 0.84	82.97 ± 1.13	99.65 ± 0.16	86.46 ± 0.86

TABLE 6. Activity recognition performance (mean ± standard deviation) obtained for ML models of the UP-FALL in [27]. Best results are highlighted in bold.

Modality	Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F_1-Score (%)
CAM	RF	32.33 ± 0.90	14.45 ± 1.07	14.48 ± 0.82	92.91 ± 0.09	14.38 ± 0.89
	SVM	**34.40 ± 0.67**	13.81 ± 0.22	14.30 ± 0.31	92.97 ± 0.06	13.83 ± 0.27
	MLP	27.08 ± 2.03	8.59 ± 1.69	10.59 ± 0.38	92.21 ± 0.09	7.31 ± 0.82
	KNN	34.03 ± 1.11	**15.32 ± 0.73**	**15.54 ± 0.57**	**93.09 ± 0.11**	**15.19 ± 0.52**
	Average	31.96 ± 1.17	13.04 ± 0.92	13.73 ± 0.52	92.79 ± 0.08	12.67 ± 0.62

single frame. However, when using classifiers that provide an estimation of probability for each class for each detection, it might be possible to identify and possibly ignore these ambiguous cases. Also, sequence-based classification such as LSTM (Long Short Term Memory) might be able to resolve these more difficult cases.

For activity detection, in comparison with the results obtained by [27] (Table 6), better results were obtained in all the metrics by all four classifiers. An improvement in the metrics accuracy of between 65.25% and 71.85% has been obtained. As this work uses the same four classifiers, it can be inferred that the use of skeletons has led to these significant improvements.

It is important to highlight that the activity recognition method proposed here delivers state-of-the-art results for UP-FALL, even surpassing the results using Convolutional Neural Networks (CNN) in [27] (accuracy = 95.1% and F_1-Score = 71.2%), with which they obtain the best results of their paper. Fig 6 compares the confusion matrices of our best model and the CNN in [27].

Also, the method proposed exceeds the results of the Challenge UP:Multimodal Fall Detection competition [44]. Three of the four ML-models presented here exceed the results of the winning entry [43] in that competition for all the performance metrics. That work combines multiple portable inertial sensors, accelerometer and gyroscope, to recognize activities and detect falls. Their accuracy and F_1-Score are 98.03% and 82.47%, while in the proposed approach they are 99.45% and 92.34%, respectively.

The main advantages of the proposed approach are: 1) It shows that human activity can be recognised by means of the posture of a person on a video image. The posture can be defined by a set of features (key points) that represents the main joints of the human skeleton. The approach obtains the features vector from only one frame or image, and it shows a very high performance compared to other state-of-the-art methods; 2) The feature vector (the joints of the skeleton) is easily interpretable by humans. It opens a way to perform many pre-processing steps before actually using the classification model. For instance, if there are many people in the scene, the classification model could be applied to only the skeletons detected with high confidence by the feature extraction method. 3) The approach can also be useful to perform multi-person activity recognition in the same scene (frame). This attribute is not discussed much in the literature, and some referenced works ([27], [30]) do not consider this scenario.

On the other hand, the main limitations of this approach are the following: 1) The results have not been contrasted with methods used for tracking. The approach does not exploit temporal consistency, although this enhancement could be implemented in the future; 2) The application is slower compared to the methods referenced above, because it consists of three steps: 1) the person detection implemented mainly by a CNN, 2) the feature extraction step (the human skeleton joints), and 3) the application of the classification model. These three steps make the proposed approach slower than other previously mentioned methods (e.g. [27], [30]), which use only one step.

To assess the efficiency and robustness of this approach, additional experiments have been performed with another public dataset (URFall [30]) and with other classification

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3061626, IEEE Access

FIGURE 6. Confusion matrices in activity recognition. Skeleton+RF (right), CNN [27].

TABLE 7. Results of our approach using the URFall database.

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F_1-Score (%)
URFall [30] best result	94.99	89.57	100	91.25	**94.49**
RF	99.11 ± 0.43	99.18 ± 0.59	97.53 ± 1.81	99.71 ± 0.21	98.34 ± 0.80
SVM	98.60 ± 0.30	96.50 ± 0.88	98.37 ± 0.88	98.69 ± 0.32	$\mathbf{97.42 \pm 0.60}$
MLP	90.79 ± 4.14	86.63 ± 13.60	83.06 ± 11.61	93.69 ± 8.15	83.19 ± 5.55
KNN	98.88 ± 0.31	98.41 ± 0.96	97.41 ± 1.11	99.43 ± 0.33	$\mathbf{97.90 \pm 0.60}$
AdaBoost	98.95 ± 0.31	98.42 ± 0.98	97.67 ± 1.12	99.43 ± 0.34	$\mathbf{98.04 \pm 0.59}$
XGBoost	99.26 ± 0.27	99.11 ± 0.53	98.12 ± 1.16	99.68 ± 0.19	$\mathbf{98.61 \pm 0.51}$

models (Adaboost [47] and XGBoost [48]). Table 7 shows these results. As can be seen, for the same classification models (SVN, KNN) used in [30], our approach shows better results ($F_1 - Score$: over 97%) vs. ($F_1 - Score$: over 94%). Besides, if we apply our approach with different classification models, the results are also better than [30]: for Adaboost and XGBoost the $F_1 - Score$ is over 98%.

V. CONCLUSION

In this paper, a camera-vision-based fall detection and activity recognition system with four classifier models is presented (RF, SVM, MLP and KNN). Pose estimation is proposed as a feature extraction mechanism that allows RGB images to be described by sets of human skeletons and considering the most prominent skeleton per image. The method is evaluated using a single-camera RGB modality from the UP-FALL publicly available dataset and demonstrates superior performance against other fall detection and activity recognition systems on the same data and comparable results against methods that use all the modalities of the UP-FALL dataset.

The main advantage of this work recognises human activity using a person's posture on a video image. The posture can be defined by a set of features (key points) representing the main joints of the human skeleton. The approach obtains the features vector from only one frame or image. It shows a high performance compared to other state-of-the-art methods. This approach also allows the detection of more than one person, which is not addressed by other referenced works. It could also be adapted to multi-person activity recognition.

Future work will consider developing an algorithm dealing with multi-person detection and also avoiding the confusion with the laying activity, e.g. by considering temporal methods, such as LSTM, to eliminate the problem of confusion when carrying out analysis over time, managing to identify the skeleton of interest-based on a sequence of frames.

REFERENCES

[1] A. Murad and J.-Y. Pyun, "Deep recurrent neural networks for human activity recognition," *Sensors*, vol. 17, no. 11, p. 2556, Nov. 2017.

[2] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, Mar. 2019.

[3] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li, "A review on human activity recognition using vision-based method," *J. Healthcare Eng.*, vol. 2017, Jul. 2017, Art. no. 3090343.

[4] X. Kang, B. Huang, and G. Qi, "A novel walking detection and step counting algorithm using unconstrained smartphones," *Sensors*, vol. 18, no. 1, p. 297, Jan. 2018.

[5] S. Lao, D. Wang, F. Li, and H. Zhang, "Human running detection: Benchmark and baseline," *Comput. Vis. Image Understand.*, vol. 153, pp. 143–150, Dec. 2016.

[6] B. Ammar, N. Rokbani, and A. M. Alimi, "Learning system for standing human detection," in *Proc. IEEE Int. Conf. Comput. Sci. Autom. Eng.*, vol. 4, Jun. 2011, pp. 300–304.

[7] S. Mekruksavanich, N. Hnoohom, and A. Jitpattanakul, "Smartwatch-based sitting detection with human activity recognition for office workers syndrome," in *Proc. Int. ECTI Northern Sect. Conf. Electr., Electron., Comput. Telecommun. Eng. (ECTI-NCON)*, Feb. 2018, pp. 160–164.

[8] D. Banerjee, C. L. Daigle, B. Dong, K. Wurtz, R. C. Newberry, J. M. Siegford, and S. Biswas, "Detection of jumping and landing force in laying hens using wireless wearable sensors," *Poultry Sci.*, vol. 93, no. 11, pp. 2724–2733, Nov. 2014.

[9] B. S. Daga, A. A. Ghatol, and V. M. Thakare, "Silhouette based human fall detection using multimodal classifiers for content based video retrieval systems," in *Proc. Int. Conf. Intell. Comput., Instrum. Control Technol. (ICICICT)*, Jul. 2017, pp. 1409–1416.

[10] P. Bet, P. C. Castro, and M. A. Ponti, "Fall detection and fall risk assessment in older person using wearable sensors: A systematic review," *Int. J. Med. Informat.*, vol. 130, Oct. 2019, Art. no. 103946.

[11] J. A. Stevens, P. Corso, E. Finkelstein, and T. Miller, "The costs of fatal and non-fatal falls among older adults," *Injury Prevention*, vol. 12, no. 5, pp. 290–295, 2006.

[12] S. Khojasteh, J. Villar, C. Chira, V. González, and E. de la Cal, "Improving fall detection using an on-wrist wearable accelerometer," *Sensors*, vol. 18, no. 5, p. 1350, Apr. 2018.

[13] Z. Liu, Y. Cao, L. Cui, J. Song, and G. Zhao, "A benchmark database and baseline evaluation for fall detection based on wearable sensors for the Internet of medical things platform," *IEEE Access*, vol. 6, pp. 51286–51296, 2018.

[14] S. B. Kwon, J.-H. Park, C. Kwon, H. J. Kong, J. Y. Hwang, and H. C. Kim, "An energy-efficient algorithm for classification of fall types using a wearable sensor," *IEEE Access*, vol. 7, pp. 31321–31329, 2019.

[15] A. Tuan Nghiem, E. Auvinet, and J. Meunier, "Head detection using kinect camera and its application to fall detection," in *Proc. 11th Int. Conf. Inf. Sci., Signal Process. Their Appl. (ISSPA)*, Jul. 2012, pp. 164–169.

[16] Y. Nizam, M. Mohd, and M. Jamil, "Development of a user-adaptable human fall detection based on fall risk levels using depth sensor," *Sensors*, vol. 18, no. 7, p. 2260, Jul. 2018.

[17] L. Panahi and V. Ghods, "Human fall detection using machine vision techniques on RGB–D images," *Biomed. Signal Process. Control*, vol. 44, pp. 146–153, Jul. 2018.

[18] C. Zhangjie and Y. Wang, "Infrared–ultrasonic sensor fusion for support vector machine–based fall detection," *J. Intell. Mater. Syst. Struct.*, vol. 29, no. 9, pp. 2027–2039, 2018.

[19] K. de Miguel, A. Brunete, M. Hernando, and E. Gambao, "Home camera-based fall detection system for the elderly," *Sensors*, vol. 17, no. 12, p. 2864, Dec. 2017.

[20] S. C. Agrawal, R. K. Tripathi, and A. S. Jalal, "Human-fall detection from an indoor video surveillance," in *Proc. 8th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2017, pp. 1–5.

[21] S. Choi and S. Youm, "A study on a fall detection monitoring system for falling elderly using open source hardware," *Multimedia Tools Appl.*, vol. 78, pp. 28423–28434, Dec. 2019.

[22] E. Cippitelli, F. Fioranelli, E. Gambi, and S. Spinsante, "Radar and RGB-depth sensors for fall detection: A review," *IEEE Sensors J.*, vol. 17, no. 12, pp. 3585–3604, Jun. 2017.

[23] C.-L. Liu, C.-H. Lee, and P.-M. Lin, "A fall detection system using k-nearest neighbor classifier," *Expert Syst. Appl.*, vol. 37, no. 10, pp. 7174–7181, Oct. 2010.

[24] T. Theodoridis, V. Solachidis, N. Vretos, and P. P. Daras, "Human fall detection from acceleration measurements using a recurrent neural network," in *Proc. Int. Conf. Biomed. Health Inform.* Cham, Switzerland: Springer, 2017, pp. 145–149.

[25] A. Núñez-Marcos, G. Azkune, and I. Arganda-Carreras, "Vision-based fall detection with convolutional neural networks," *Wireless Commun. Mobile Comput.*, vol. 2017, Dec. 2017, Art. no. 9474806.

[26] O. Aziz, J. Klenk, L. Schwickert, L. Chiari, C. Becker, E. J. Park, G. Mori, and S. N. Robinovitch, "Validation of accuracy of SVM-based fall detection system using real-world fall and non-fall datasets," *PLoS ONE*, vol. 12, no. 7, Jul. 2017, Art. no. e0180318.

[27] L. Martínez-Villaseñor, H. Ponce, J. Brieva, E. Moya-Albor, J. Núñez-Martínez, and C. Peñafort-Asturiano, "UP-fall detection dataset: A multimodal approach," *Sensors*, vol. 19, no. 9, p. 1988, Apr. 2019.

[28] X. Ma, H. Wang, B. Xue, M. Zhou, B. Ji, and Y. Li, "Depth-based human fall detection via shape features and improved extreme learning machine," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 6, pp. 1915–1922, Nov. 2014.

[29] Z. Zhang, C. Conly, and V. Athitsos, "Evaluating depth-based computer vision methods for fall detection under occlusions," in *Advances in Visual Computing*. Cham, Switzerland: Springer, 2014, pp. 196–207.

[30] B. Kwolek and M. Kepski, "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Comput. Methods Programs Biomed.*, vol. 117, no. 3, pp. 489–501, Dec. 2014.

[31] A. Dib and F. Charpillet, "Pose estimation for a partially observable human body from RGB-D cameras," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 4915–4922.

[32] S. Gasparrini, E. Cippitelli, E. Gambi, S. Spinsante, J. Wåhslén, I. Orhan, and T. Lindh, "Proposal and experimental evaluation of fall detection solution based on wearable and depth data fusion," in *ICT Innovations*, S. Loshkovska and S. Koceski, Eds. Cham, Switzerland: Springer, 2016, pp. 99–108.

[33] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding," 2017, *arXiv:1703.07475*. [Online]. Available: http://arxiv.org/abs/1703.07475

[34] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.

[35] P. S. Sase and S. H. Bhandari, "Human fall detection using depth videos," in *Proc. 5th Int. Conf. Signal Process. Integr. Netw. (SPIN)*, Feb. 2018, pp. 546–549.

[36] F. Merrouche and N. Baha, "Depth camera based fall detection using human shape and movement," in *Proc. IEEE Int. Conf. Signal Image Process. (ICSIP)*, Aug. 2016, pp. 586–590.

[37] S. Jeong, S. Kang, and I. Chun, "Human-skeleton based fall-detection method using LSTM for manufacturing industries," in *Proc. 34th Int. Tech. Conf. Circuits/Syst., Comput. Commun. (ITC-CSCC)*, Jun. 2019, pp. 1–4.

[38] B.-H. Wang, J. Yu, K. Wang, X.-Y. Bao, and K.-M. Mao, "Fall detection based on dual-channel feature integration," *IEEE Access*, vol. 8, pp. 103443–103453, 2020.

[39] H. Li, C. Li, and Y. Ding, "Fall detection based on fused saliency maps," *Multimedia Tools Appl.*, vol. 80, pp. 1883–1900, Sep. 2020.

[40] U. Iqbal, A. Milan, and J. Gall, "PoseTrack: Joint multi-person pose estimation and tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2011–2020.

[41] T.-H. Tsai and C.-W. Hsu, "Implementation of fall detection system based on 3D skeleton for deep learning technique," *IEEE Access*, vol. 7, pp. 153049–153059, 2019.

[42] Q. Xu, G. Huang, M. Yu, and Y. Guo, "Fall prediction based on key points of human bones," *Phys. A, Stat. Mech. Appl.*, vol. 540, Feb. 2020, Art. no. 123205.

[43] H. Gjoreski, S. Stankoski, I. Kiprijanovska, A. Nikolovska, N. Mladenovska, M. Trajanoska, B. Velichkovska, M. Gjoreski, M. Luštrek, and M. Gams, *Wearable Sensors Data-Fusion and Machine-Learning Method for Fall Detection and Activity Recognition*. Cham, Switzerland: Springer, 2020.

[44] H. Ponce, M. de Lourdes Martínez-Villaseñor, J. Brieva, and E. Moya-Albor, *Challenges and Trends in Multimodal Fall Detection for Healthcare*. Cham, Switzerland: Springer, 2020.

[45] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2334–2343.

[46] A. Collado-Villaverde, M. Cobos, P. Muñoz, and D. F. Barrero, "A simulator to support machine learning-based wearable fall detection systems," *Electronics*, vol. 9, no. 11, p. 1831, Nov. 2020.

[47] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[48] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

HEILYM RAMIREZ was born in Bogotá, Colombia, in 1993. She received the degree in electronic engineering from the Distrital University "Francisco Jose de Caldas" (UD), Bogotá, in 2016. Her current research interest includes machine learning.

SERGIO A. VELASTIN (Senior Member, IEEE) received the B.Sc. and M.Sc. (Research) degrees in electronics and the Ph.D. degree from The University of Manchester, Manchester, U.K., in 1978, 1979, and 1982, respectively, for research on vision systems for pedestrian tracking and road-traffic analysis. He worked with industrial research and development before joining King's College London, U.K., in 1991, and then Kingston University London, where he became the Director of the Digital Imaging Research Centre and a Full Professor of Applied Computer Vision. In 2015, he moved to the University Carlos III of Madrid, Spain, where he was a Marie Curie Professor. He is currently a Visiting Professor with the Universidad Carlos III, Madrid, and the Queen Mary University of London, U.K.

IGNACIO MEZA was born in La Serena, Chile, in 1995. He is currently pursuing the degree in electrical engineering with the Pontificia Universidad Católica de Valparaíso. His research interests include machine learning, computer vision, and quantum machine learning.

ERNESTO FABREGAS received the B.S. degree in automation control and the M.S. degree in digital systems from the Universidad Tecnológica de La Habana "José Antonio Echeverría", Cujae, Cuba, in 2004 and 2008, respectively, and the Ph.D. degree in computer science from the Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain, in 2013. From 2015 to 2019, he was a Postdoctoral Fellow with UNED. Since 2019, he has been an Assistant Professor with UNED. His current research interests include control of multi-agent systems, machine learning, images processing, mobile robot control, remote laboratories, and engineering education.

DIMITRIOS MAKRIS received the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1999, and the Ph.D. degree in computer vision from City University, London, U.K., in 2004. He is currently a Professor with the School of Computer Science and Mathematics, Kingston University, London, U.K. His research interests include computer vision, machine learning, and particularly video analysis and human motion analysis. He is a member of the British Machine Vision Association. He is currently the Chair of the IET Vision and Imaging Technical Network.

GONZALO FARIAS received the B.S. degree in computer science from the De La Frontera University, Temuco, Chile, in 2001, the Ph.D. degree in control engineering from the National University of Distance Education (UNED), in 2010, and the Ph.D. degree in computer science from the Complutense University of Madrid (UCM), Madrid, Spain, in 2013. Since 2012, he has been an Assistant Professor with the Electrical Engineering School, Pontificia Universidad Católica de Valparaíso (PUCV), Chile. His current research interests include machine learning, pattern recognition, simulation and control of dynamic systems, and engineering education.