M. Mario, "Human Activity Recognition Based on Single Sensor Square HV Acceleration Images and Convolutional Neural Networks," in IEEE Sensors Journal, vol. 19, no. 4, pp. 1487-1498, 15 Feb.15, 2019

DOI: https://dx.doi.org/10.1109/JSEN.2018.2882943

Human Activity Recognition based on Single Sensor Square HV Acceleration Images and Convolutional Neural Networks

Munoz-Organero Mario, *Member, IEEE*

***Abstract*—Human Activity Recognition (HAR) provides the context for many user-centered personal recommender systems in areas such as healthcare, sports, lifelong learning or home automation. Based on different types of sensors (either camera based, environmental sensors or wearable and mobile sensors) user related data provides the basis to extract movement related features from which the activity that the user is performing can be assessed. Among the different types of sensors, wearable sensors provide a user convenient, non-intrusive, always available alternative that has gained special attention for HAR. Wearable sensors will be a relevant part of the Internet of Things. This paper presents a novel mechanism to detect which particular activity a user is performing based on the data from a single tri-axial accelerometer. A Convolutional Neural Network (CNN) is used in order to automatically extract the most relevant features to characterize acceleration patterns with inter-activity discrimination capacity. The user anchored coordinate system generating the data from the accelerometer sensor is transformed into a georeferenced coordinate system in order to estimate the horizontal and vertical acceleration components. A sliding window with 50% overlap is used to extract 5 seconds of acceleration data from which a square horizontal-vertical (HV) acceleration image is computed. Both monochrome and colored images are generated either by adding the influence of the time evolution of the acceleration series or not in the generated image. The results for both p-fold cross-validation and leave on out (L1O) approaches are presented using a public dataset. The results outperform by around 8% those obtained by the authors of the dataset in the case of using a p-fold cross-validation.**

***Index Terms*—Human Activity Recognition in an Internet of Things, Convolutional Neural Networks, Acceleration Images, Accelerometer sensors.**

I. INTRODUCTION

THE information provided by user-worn sensors such as accelerometers and gyroscopes has proven to be a valid candidate in order to estimate the activity that the user is performing [1-3]. Common Human Activity Recognition (HAR) architectures tend to use a cascade of computational steps in order to transform the recorded sensor data into a corresponding activity. Raw sensor data is first pre-processed and converted into movement related features which are then used to feed a classifier which is trained in a supervised way using some pre-tagged (labeled) data. Depending on the mechanism used to define the movement related features the previous research could be classified into hand-crafted mechanisms or deep-learning based automatic feature characterization and extraction methods. Depending on the type of classifier used to convert the values of a particular set of features into a related user activity, several classifications could be used comprising datamining techniques such as decision trees, Support Vector Machines (SVM) or K-Nearest Neighbor (KNN), probabilistic models with hidden latent variables such as Hidden Markov Models (HMM), Fuzzy Logic based systems such as Fuzzy Inference Systems (FIS), Bayesian probabilistic models (either parametric and non-parametric) and Deep Learning (DL) based methods such as Deep Belief Networks (DBN) and Convolutional Neural Networks (CNN). Early studies in related literature such as [4] are based on the use of hand-crafted features over data from a single accelerometer using classifiers such as decision Trees (C4.5), K-nearest neighbors (KNN), Support Vector Machines (SVM), Naive Bayes and even ensemble methods such as Boosting and Bagging. The combining of several sensors in order to improve the classification results is also present in some early studies such as [5] in which acceleration data from 5 bi-axial accelerometers and several classifiers (nearest neighbor, C4.5 decision tree, and naive Bayes classifiers) were used to recognize 20 different activities. Early studies such as [4][5] used hand-crafted features such as the mean temporal value in a window of raw data, the energy of the raw signal and some frequency-domain parameters such as entropy. The authors in [2] categorized the hand-crafted features into two major categories, statistical and structural based. Statistical methods, such as the Fourier transform and the Wavelet transform, use quantitative characteristics of the data to extract features, whereas structural approaches take into account the interrelationship among data [2]. The major contributions to Human Activity Recognition (HAR) based on hand-crafted features and pre-deep-learning classification methods have been captured in several survey studies such as [1-3].

This work was supported by the "ANALYTICS USING SENSOR DATA FOR FLATCITY" project TIN2016-77158-C4-1-R (MINECO/ ERDF, EU) funded by the Spanish Agencia Estatal de Investigación (AEI) and the European Regional Development Fund (ERDF)."

M. Munoz-Organero is with the Department of Telematics Engineering and UC3M-BS Institute of Financial Big Data, Universidad Carlos III de Madrid, 28911 Leganes, Madrid, Spain (e-mail: munozm@it.uc3m.es).

The use of Deep Learning (DL) techniques for Human Activity recognition provides mechanisms to automatically extract suitable features from raw sensor data which are learned in the algorithm training phase. Among the different DL algorithms, Convolutional Neural Networks have shown especial interest in the recent years [6]. By stacking together the time series of either a single multi-axis sensor or the time series of several separated independent sensors, input images have been generated and used to train and validate different CNN structures providing better results comparted to previous non-DL methods. A review of the major research studies using CNNs for HAR will be presented in the next section.

HAR algorithms can be trained for a single user or multiple users activity recognition. Single user personal models are trained with data from a single user and then applied to detect activity related information only to new data from that user while multi-user models are either trained with data form a set of users and applied for activity recognition to the data of a different set of users, or trained with a sub-set of data from the same set of users, whose activities are going to be estimated, based on a different sub-set of data from the same users. Personal models tend to perform dramatically better than the multi-user models, even when trained from only a few minute worth of data [7]. However, it is not always possible to train the algorithms with data from each particular user and multi-user generalization models are normally required. When training the HAR algorithm with multiple user data, two major approaches are used in order to validate the accuracy of results. A Leave one out approach (L1O) uses the data from all the users except one to train the algorithms and the data from the left apart user to validate the performance of the algorithm for human activity recognition. The process is repeated by leaving apart each user and the average of results is provided. A p-fold cross-validation approach, uses a random percentage of the data from all the users to train the HAR algorithm and the rest of the data to validate results. The process is repeated until all the data has been used once for validation and the average results are provided. As reported in previous studies, when L1O cross validation is used, the classification rates are normally lower than when using a p-fold cross-validation approach [8]. The way each particular user performs each activity to be detected tends to show a bigger correlation with previous executions of the same activity by the same user than with other users performing the same activity. The generalization of results in a L1O approach will depend on having similar users in the training set to the one left apart. In order to improve the classification accuracy, cross-subject models such as clustering the users into similar groups in order to apply models trained to similar users for HAR and using model-based adaptation approaches such as adjustable fuzzy clustering, have been explored in previous studies such as [9].

The performance of acceleration data based HAR algorithms is also influenced by the location where the accelerometer sensor is attached to the body. The impact of the position of the sensor to recognize activities has been studied in previous research papers such as [10] and [11]. Some locations perform better for a particular activity. When using cross-subject models, the authors in [11] identified the waist as the most suitable device location at which the acceleration patterns for the same activity across several people are most similar.

This paper uses both p-fold cross-validation and L1O approaches to validate a new method to feed a CNN with acceleration images in order to recognize human activities using a public dataset. The activities to be recognized are climbing stairs down and up, jumping, running/jogging and walking. A single tri-axial accelerometer located in the waist region is used. The results outperform previous techniques using the same dataset by around 8%. The major novelty introduced in this paper focuses on the way in which the acceleration images are computed for the particular case of HAR and comprises 3 major aspects:

1) The use of 28 by 28 square images. Previous research studies computing acceleration images from wearable sensors for HAR have focused on rectangular images.
2) Transforming the sensed acceleration values in a sensor anchored coordinate system into vertical (perpendicular to the ground) and horizontal (parallel to the ground) acceleration values which are independent of the sensor orientation and easier to interpret.
3) Capturing the time information in the acceleration data as colors in the generated images.

The rest of the paper is organized as follows. Section 2 captures a review of previous research studies using CNNs for Human Activity Recognition (HAR). The major previous findings will be used as the basis in the current paper. Section 3 is dedicated to describing the proposed algorithm which is based on the computation of square acceleration images using a geo-referenced coordinate system, either using 2D (monochrome) or 3D (colored) images. Section 4 describes the database used as well as the validation techniques that will be considered in order to estimate the accuracy of the proposed algorithm. Section 5 describes the major results and section 6 captures the main conclusions.

II. Convolutional Neural Networks for HAR

The use of Deep Learning (DL) algorithms to automatically learn the most relevant features to recognize human activities has recently provide very promising results both in the case of a single accelerometer [12] or when multiple accelerometer and gyroscope sensors are used [13]. Among the different DL methods, CNN based on rectangular acceleration images have been preferred [12]. This section provides a review of previous research studies using CNN for HAR as the basis for the method proposed in this paper.

A Convolutional Neural Network stacks an arbitrary long sequence of convolutional and pooling layers which are ended by adding one or several fully connected layers [14]. Each pooling layer uses a set of convolutional kernels which are automatically learned based on the training data. The input of the CNN can be 2D (monochrome) or 3D (colored) images. When using a CNN for HAR using acceleration data, a mechanism to generate 2D or 3D images from acceleration time series has to be provided. Previous research studies in HAR have generated 2D acceleration images by stacking together the time series of a single or several multi-axis accelerometers in a few seconds window of raw data. The acceleration values over time in each sensor axis constitute one dimension of the generated image. The sensed values for each different axis and sensor at the same time provide the information for the other dimension in the image. Even if the window used only comprises a few seconds of acceleration data, the sampling rates used to capture the required information to describe each activity, generate more data in the time dimension as compared to the number of sensors or axes available. Therefore, rectangular acceleration images are normally used in order to feed the CNN architecture. Rectangular convolutional kernels have also been used in order to adapt the information extraction process to the generated 2D rectangular acceleration images.

The authors in [12] proposed a 2D image generation process based on stacking together 8 acceleration related time series obtained from a single tri-axial accelerometer. Apart from the 3 raw acceleration components, the authors computed the movement related acceleration by estimating and compensating the gravity component and they also estimated and used the pitch and roll for the sensor orientation. A CNN was able to learn common acceleration patterns using a simple architecture even better than previous papers based on more complex architectures but only using the raw acceleration data showing that the way in which the acceleration image is computed plays a very important role in the results that can be achieved. The authors used the WHARF dataset [15] and the experimental results showed that the proposed method outperformed previous research studies. The accuracy obtained in this case for the WHARF dataset was 79.31%.

The authors in [13] used a CNN applied to multiple accelerometer and gyroscope sensors. 2D convolution and pooling operations were applied to a 2D acceleration image based on the multivariate time series, outperforming 1D techniques applied to a single series at a time. The authors used three accelerometers (in the chest, right wrist and left ankle) and two gyroscopes (in the right wrist and left ankle) sensors and used the Mhealth dataset [16]. The authors only performed intra-subject classification and the accuracy obtained was comprised between an 84% in the worst case and 97% for the best user.

The authors in [17] generated 2D images by using different axes from different sensors padded with ceros so that different types of sensors did not overlap when using the convolutional kernels. They applied intra-subject validation techniques to 2 different datasets. The authors achieved a 98.3% accuracy on the dataset in [18] and a 97.9% on the dataset in [19].

Most Inertial Measurement Units (IMU) in wearable devices include both an accelerometer and a gyroscope. The authors in [20], combined the information of the accelerometer and gyroscope from a single IMU in order to create sensor 2D images to feed a CNN. A total of 9 signals were used: the 3 components from the raw acceleration data, the 3 components from the raw gyroscope data and the 3 components from the estimated movement related acceleration (in which the gravity component had been removed). The size of the images was scaled to 36 by 68 by repeating 1 second samples of the 9 signals. The authors used 3 different datasets in order to validate results. They used part of the data from all the users to train the CNN and the rest of the data to validate their approach. Using the dataset in [21] they were able to achieve a 95.18% accuracy. They got a 97.01% accuracy with the dataset in [22]. Finally, the use of the dataset in [23] provided a 99.93% accuracy. A later publication by Jordao et al. [12], implementing the algorithm proposed in [20] achieved only a 70.08% accuracy when applied to the WHARF dataset [15].

Recognizing activities using a CNN over a single accelerometer will simplify the adoption of the technology by end users at the time of lowering the costs for manufacturers. Apart from the study in [12] previously mentioned, the authors in [24] applied a CNN to a single tri-axial accelerometer and studied the importance of adapting the convolutional kernel size to the characteristics of the acceleration images. The authors in [24] also used rectangular images based on stacking the 3 axes of time series provided by the tri-axial accelerometer and therefore, they also used a rectangular-shaped convolutional kernel. An ad-hoc dataset was also created to validate the algorithm using part of the data for training and the rest for validation (p-fold cross-validation). They obtained an average accuracy of 93.8% on their dataset. The authors in [12] applied the method in [24] to the WHARF dataset and obtained a 72.29% accuracy.

From the previous results in using CNN for HAR based on acceleration data, the way in which the acceleration images are computed have showed to be even more relevant than the architecture of the network itself. Previous studies have focused on 2D rectangular acceleration images in which the time evolution of the raw sensor signals was already one of the dimensions in the generated images. The first novelty of this paper is analyzing the benefits of generating square acceleration images and square convolutional kernels for HAR based on a single tri-axial accelerometer. Moreover, a second novelty as compared to previous studies using acceleration images to feed CNN architectures for HAR is that the sensor anchored coordinate system, attached to the user's body and generating the data, is transformed into a georeferenced coordinate system in order to estimate the horizontal and vertical acceleration components, compensating for different orientations of the sensor when attached to the body and providing easier to interpret data (up-down versus parallel to the ground movements). Finally, a third novelty introduced in the paper is that the time information is added to the Horizontal-Vertical (HV) square acceleration images by using 2 color components.

III. Proposed Algorithm

The proposed algorithm applies a Convolutional Neural Network (CNN) to the information obtained from a single tri-axial accelerometer. Summarizing previous research studies, as captured in the previous section, there are two major aspects that can be optimized in order to improve the performance of a CNN for Human Activity Recognition (HAR):

1) The architecture of the network and its parameters such as the number of convolutional layers and the size of the convolutional kernels

2) The way in which the acceleration information is presented to the CNN in terms of acceleration images

Regarding the architecture of the network, the research in [12] showed that simpler architectures could out-perform preciously proposed more complex architectures by optimizing the way in which the information is presented to the network. This paper follows this recommendation by using a similar network architecture as the one in [12] with 3 convolutional layers and a final fully connected layer and focuses on providing a new way for generating acceleration images in order to achieve better accuracy. The CNN is trained to recognize 5 different activities: climbing stairs down and up, jumping, running/jogging and walking. The data from the 3-axial accelerometer is used to compute either 2D (monochrome) or 3D (colored) square HV acceleration images based on a 5 second sliding window with a 50% overlap. Using 5 second windows for segmenting raw data for HAR has been widely used in previous research studies such as [12]. The authors in [11] state that the window size depends on the kind of activities which should be recognized considering sizes of a few seconds with 50% overlap. The idea is that each window contains enough information for the learning algorithm to have access to an entire cycle in periodic movements such as the ones considered in the current paper. The size of the window will therefore constrain the minimum speed at which movements are repeated inside the activities to be detected. The images are fed into the CNN configured in Matlab [25] with the following parameters:

```
layers = [
   imageInputLayer([28 28 “1 or 2”])
   convolution2dLayer(3,8,'Padding',1)
   batchNormalizationLayer
   reluLayer
   maxPooling2dLayer(2,'Stride',2)
   convolution2dLayer(3,16,'Padding',1)
   batchNormalizationLayer
   reluLayer
   maxPooling2dLayer(2,'Stride',2)
   convolution2dLayer(3,32,'Padding',1)
   batchNormalizationLayer
   reluLayer
   fullyConnectedLayer(5)
   softmaxLayer
   classificationLayer];
```

The graphical architecture for the CNN is captured in figure 1. The size of 28 by 28 for the input images to the CNN has been widely used in previous research studies for image recognition and in benchmark image databases such as the MNIST database of handwritten digits [26]. Monochrome images will be represented as 28 by 28 by 1 matrices while colored images will use 28 by 28 by 2 matrices as described later in this section. The filters in the CNN, for all the convolutional layers, are 3 by 3 by k matrices (being k the number of channels in the input images to each convolutional layer). At the output of each convolutional layer, a padding of “ceros” of size 1 is added to all the edges of the output image (the convolutional operation does not reduce the size of the vertical and horizontal dimensions). In order to improve the convergence in the training of the CNN, a normalization operation is added after each convolutional layer and before the non-linear ReLU layer. A final max polling layer is used at the end of the first 2 convolutional layers to reduce the vertical and horizontal sizes for the output images by a factor of 2. The number of channels is increased by a factor of 2 in each consecutive convolutional layer as the sizes of the horizontal and vertical dimensions are reduced by a similar factor of 2 as previously said. A final fully connected layer and a softmax layer assign a probability for each input image to one of the 5 classes/activities (climbing stairs down and up, jumping, running/jogging and walking). The default parameters in Matlab have been used for the weight and bias learning rates in the training of the CNN.

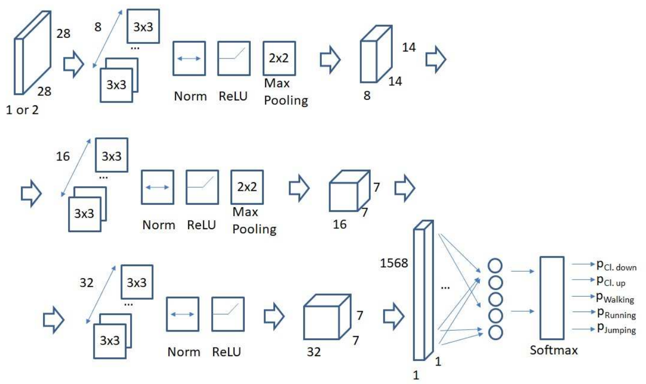

Fig. 1. Architecture for the CNN

The output layer assigns probabilities to each image for each class (activity) based on the *softmax* layer and selects the activity with a higher probability. A minimum threshold could be used in order to assign a particular window of acceleration data to one of the activities or mark that window as “unknown”. Using validation datasets in which target activities are performed mixed with inter-activity transitions (or intra-activity resting segments) could benefit from this “unknown” class which should detect such regions. For the current paper however, each data window has been assigned to one of the target classes (activities) in order to compare results with those of the researchers that recorded the information in the dataset.

Together with the architecture for the network, a mechanism to generate input images to feed the CNN from acceleration data is needed. This section describes the proposed mechanisms in order to generate the HV square images that will be used to feed the CNN architecture in 2 sub-sections (one for monochrome and another for colored images). The idea is to estimate the movement related acceleration and its projections in the vertical (up and down movement caused accelerations) and horizontal (parallel to the ground) components. Each data point in the horizontal and vertical acceleration time series will be mapped into a horizontal-vertical (HV) 28 by 28 image in order to draw a monochrome image. The time related information can be added into the image by coloring it (3D acceleration images).

A. Monochrome HV acceleration images

In this sub-section, the data from a single 3-axial accelerometer is used to compute 2D acceleration images based on a 5 second sliding window with 50% overlap. The raw acceleration data in the sensor coordinate systems sampled at 50*Hz* is transformed into a geo-referenced 28 * 28 HV square image using the following algorithm:

1) The 3-axial accelerometer generates periodic acceleration samples in the sensor coordinate system: $\overrightarrow{a_t} = (a_{tx}, a_{ty}, a_{tz})$
2) The acceleration vector caused by the gravity force in the sensor coordinate system is estimated using a low pass filter using the following equation: $\overrightarrow{g'} = \frac{\sum_{5\,seconds}(a_{tx}, a_{ty}, a_{tz})}{5*f}$ where f is sampling frequency (50 *Hz* for the database used in this paper).
3) The acceleration caused by the human movement (gravity free acceleration) is estimated according to the following equation: $\overrightarrow{a_m} = \overrightarrow{a_t} - \overrightarrow{g'}$
4) The computed $\overrightarrow{a_m}$ vector is then divided into its vertical and horizontal acceleration components following the equation: $\overrightarrow{a_v} = \frac{\overrightarrow{a_m} \cdot \overrightarrow{g'}}{\|\overrightarrow{g'}\|^2} \overrightarrow{g'}$ and $\overrightarrow{a_h} = \overrightarrow{a_m} - \overrightarrow{a_v}$
5) A 5 second sliding window (250 samples for a 50 *Hz* sampling rate) is used to generate an acceleration image using the following pseudo-code:

Let img be a 28*28 image initialized to all zeros
for i=1 to 250

indexv= $floor\left(15 + 6.9 * min\left(2, \frac{\|\overrightarrow{a(\iota)_v}\|}{\|\overrightarrow{g(\iota)'}\|}\right)\right)$

indexh= $floor\left(1 + 13.9 * min\left(2, \frac{\|\overrightarrow{a(\iota)_h}\|}{\|\overrightarrow{g(\iota)'}\|}\right)\right)$

img(indexh,indexv) = img(indexh,indexv) + 1

6) Store the acceleration image, move the window 125 samples (2.5 seconds) and continue with step 5.
7) The acceleration images are computed for all participants. Part of the data will be used for training of the CNN and part for validation as described in the following sub-sections.

The resulting images are 28 by 28 and use a geo-referenced coordinate system based on estimating the movement related acceleration perpendicular to the ground (in the direction of the gravity force) and the acceleration component parallel to the ground. The accelerometer sensor generates acceleration samples in a sensor anchored coordinate system. Since the sensor is attached to the participant's body, the sensor coordinate system will move with the user. A geo-referenced coordinate system capturing the acceleration vertical and parallel to the ground is used so that the acceleration samples generate acceleration images independent of the sensor orientation. The acceleration samples captured by the accelerometer sensor include both the gravity and movement related acceleration components (step 1 above). The estimation of the gravity force allows both to compensate the acceleration samples in order to obtain the movement related acceleration and to obtain the projections of the acceleration samples into the vertical and perpendicular to the ground components (steps 2 to 4 above). The gravity component is estimated by averaging the acceleration samples over a 5 second window (step 2 above). Similar approaches both to estimate and compensate the gravity related component from raw acceleration data have been previously proposed in research studies such as [28, 29]. As previously captured, the size of 28 by 28 for square images feeding a CNN has been widely used in previous research studies for image recognition and in benchmark image databases such as the MNIST database of handwritten digits [26]. The acceleration images are generated using the estimated vertical and horizontal accelerations trying to minimize the effect of the different placements of the sensors for each participant and the relative to the body sensor moves caused by a loose sensor strap.

B. Colored HV acceleration images

2D (monochrome) acceleration images take into account the statistical information of each acceleration sample as if the different samples were independent. In this sub-section a coloring mechanism is proposed in order to add the time dependency among different acceleration samples. The proposed algorithm uses 2 different colors (C1 and C2) and creates a weighted projection of acceleration samples as computed in the monochrome image to both colors which takes into account the time component (and time dependencies for the acceleration data samples). A 3D (colored) image is therefore created. Figure 1 shows the option to feed the CNN with 28 by 28 by 2 matrices, being the last dimension the number of colors used. The proposed algorithm is described in the following steps:

1) The 3-axial accelerometer generates periodic acceleration samples in the sensor coordinate system: $\overrightarrow{a_t} = (a_{tx}, a_{ty}, a_{tz})$
2) The acceleration vector caused by the gravity force in the sensor coordinate system is estimated using a low pass filter using the following equation: $\overrightarrow{g'} = \frac{\sum_{5\,seconds}(a_{tx}, a_{ty}, a_{tz})}{5*f}$ where f

is sampling frequency (50 *Hz* for the database used in this paper).

3) The acceleration caused by the human movement (gravity free acceleration) is estimated according to the following equation: $\overrightarrow{a_m} = \overrightarrow{a_t} - \overrightarrow{g'}$

4) The computed $\overrightarrow{a_m}$ vector is then divided into its vertical and horizontal acceleration components following the equation: $\overrightarrow{a_v} = \frac{\overrightarrow{a_m} \cdot \overrightarrow{g'}}{\left\|\overrightarrow{g'}\right\|^2} \overrightarrow{g'}$ and $\overrightarrow{a_h} = \overrightarrow{a_m} - \overrightarrow{a_v}$

5) A 5 second sliding window (250 samples for a 50 *Hz* sampling rate) is used to generate an acceleration image using the following pseudo-code:

Let img be a 28*28*2 image initialized to all zeros
for i=1 to 250

$$\text{indexv}=floor\left(15 + 6.9 * min\left(2, \frac{\left\|\overrightarrow{a(\iota)_v}\right\|}{\left\|\overrightarrow{g(\iota)'}\right\|}\right)\right)$$

$$\text{indexh}=floor\left(1 + 13.9 * min\left(2, \frac{\left\|\overrightarrow{a(\iota)_h}\right\|}{\left\|\overrightarrow{g(\iota)'}\right\|}\right)\right)$$

$$C1 \rightarrow img(indexh, indexv, 1) = img(indexh, indexv, 1) + 1 - \frac{(i-1)}{250}$$

$$C2 \rightarrow img(indexh, indexv, 2) = img(indexh, indexv, 2) + \frac{i}{250}$$

6) Store the acceleration image, move the window 125 samples (2.5 seconds) and continue with step 5.

7) The acceleration images are computed for all participants. Part of the data will be used for training of the CNN and part for validation as described in the following sub-sections.

The coloring mechanism is designed by adding a different weight to each color depending on the time information. Color C1 gets a bigger weight at the beginning of the acceleration time series and the weight decreases linearly. C2 complements C1 and gets a bigger weight by the end of the time window.

IV. Dataset And Validation Methods

In order to validate the results of the proposed algorithm, a real world dataset from 15 participants for 8 common activities where they carried 7 wearable devices in different on-body positions [27] has been used. The dataset includes the acceleration data of 8 different human activities, 5 dynamic and 3 static activities: climbing stairs down, climbing stairs up, jumping, lying, standing, sitting, running/jogging, and walking of fifteen subjects (age 31.9±12.4, height 173.1±6.9, weight 74.1±13.8, eight males and seven females). For each activity, the dataset contains the simultaneously recorded data from 7 different sensor sets placed on the chest, the forearm, the head, the shin, the thigh, the upper arm, and the waist. Each activity was performed by each participant for about 10 minutes except for the jumping activity which was performed for a smaller period of time due to the physical exertion of the participants. The dataset comprises participants with a significant difference concerning fitness, age, weight, height, and movement behavior [11]. Moreover, participants recorded the activities in different places and environments (such as open air climbing up a mountain hill versus in building walking up stairs, running on a treadmill versus running in a city environment or in the countryside). Moreover, the recorded samples for same participants and activities include segments of data not related to the activity of interest such as walking to rest in the middle of a running activity. This heterogeneity in the underlying characteristics of recorded data in this dataset [27] constitutes an additional challenge for HAR as compared with other datasets previously used in literature such as [18-19][21-23].

From all the sensor sets in the database, only the waist located sensor set will be used since it was the one providing better inter-personal results in [11]. From all the sensors in the sensor set only the tri-axial accelerometer will be used. The use of sensor fusion techniques to improve the activity recognition is left for future studies.

In order to validate the results of applying the algorithm described in this paper to the dataset in [27] a multiuser approach has been used. A multiuser approach allows to assess the generalization of results to different users. According to the results of a previous research study by the authors of this dataset [11], both a p-fold cross-validation (p=5) and a leave one out (L1O) approaches are followed. The L1O approach assesses the generalization of results to new users not considered for the training of the algorithms. A 5-fold cross-validation approach uses 80% of the data from all the activities by all the participants to train the algorithm and the other 20% for validation (the process is repeated 5 times until all the data has been used once and only once in the validation set). Since we train the learning algorithm with data from all participants, it will be more likely that the data in the validation set is similar (in statistical terms in the feature space) to the data in the training set and the expected accuracy will be higher than in the case of using a L1O approach. As previously presented in the paper in relation to previous research studies, the accuracy of the algorithm when trained leaving out all the information of one particular individual will depend on the fact that there is a similar individual in the training set. Particular users performing activities in a peculiar or characteristic way will find a very low accuracy when using the L1O validation approach since the algorithm has not been previously trained for similar input data. Since the number of participants in the dataset is limited, a pre-clustering of users into similar groups in order to use a L1O validation approach inside each group has not been implemented. Table I and Table II show the results obtained in [11] for the same dataset [27] for the cases of using p-fold cross-validation and L1O validation approaches. The authors use the f-score measure to compare results (the same merit figure will be used in this paper to validate the proposed mechanism). L1O results show lower f-score values. Some activities such as running and jumping are better classified while activities such as climbing up and down show the worst results.

TABLE I
P-FOLD CROSS-VALIDATION RESULTS IN [11]

Class	F-measure
walking	0.86
Running	0.92
Jumping	0.97
Climbing up	0.81
Climbing down	0.79

TABLE II
L1O VALIDATION RESULTS IN [11] USING THE WAIST ACCELEROMETER ONLY

Class	F-measure
walking	0.76
Running	0.89
Jumping	0.83
Climbing up	0.65
Climbing down	0.70

V. RESULTS

This section captures the main results of applying the proposed algorithm to detect walking, running, jumping, climbing up and climbing down activities from the dataset in [27]. The results are presented for both monochrome and colored square HV acceleration images and when using both a 5-fold cross-validation and a L1O validation approach.

The CNN architecture described in section III has been used in all cases. As a difference with previous research studies, instead of composing 2D rectangular acceleration images based on stacking together the time series of either several sensors or the different axes of a single sensor, square geo-referenced images are created (as described in the previous sections). The time information can be added as a color component to the generated images. The objective is to assess the accuracy achieved when feeding the proposed CNN architecture with square Horizontal-Vertical (HV) acceleration images. The computation of the horizontal (parallel to the ground) and vertical (in the direction of the gravity force) components for the acceleration samples is based on the estimation of the acceleration caused by the gravity force. Using the averaging mechanism described in section III, the mean value for the gravity acceleration estimation from all acceleration samples obtained is 9.751 with standard deviation of 0.126.

Figures 2 to 6 show the average contour plots for the monochrome square HV acceleration images in the execution of the 5 activities by the first participant in the dataset according to the algorithms described in section III. There is a significant visual difference for the running and jumping images from the rest of activities (which is aligned to the results in [11] in which better classification results were achieved for these 2 activities). Jumping acceleration images present 2 major distinctive active regions for the up and down movements (high acceleration areas for vertical accelerations with lower horizontal acceleration values). Running acceleration images show content for both horizontal and vertical values in a more connected region (incorporating the forward movement when running as compared to vertical jumping) There are also smaller but visually distinct differences in the HV acceleration images for the walking, climbing up and down activities both in the shape of the contours and well as with other parameters such as the amplitude of the horizontal accelerations. The walking activity for participant 1 generates images in which the central contour lines are visually more circular in shape than those generated for climbing up and down (showing a more balanced behavior between vertical and horizontal accelerations when walking forwards). Climbing up tends to show a higher presence in the vertical positive acceleration region.

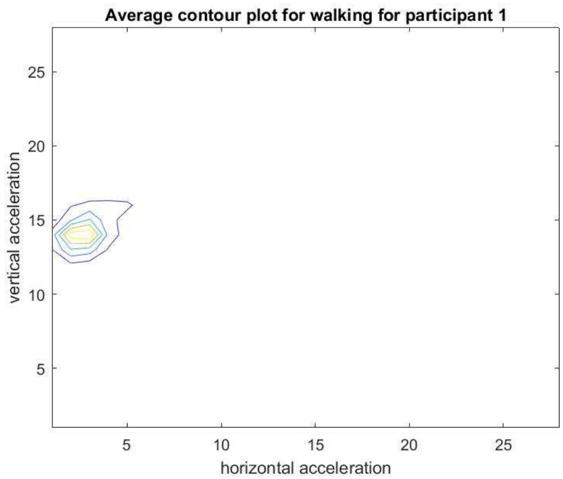

Fig. 2. Average contour plot for the walking acceleration images, participant 1

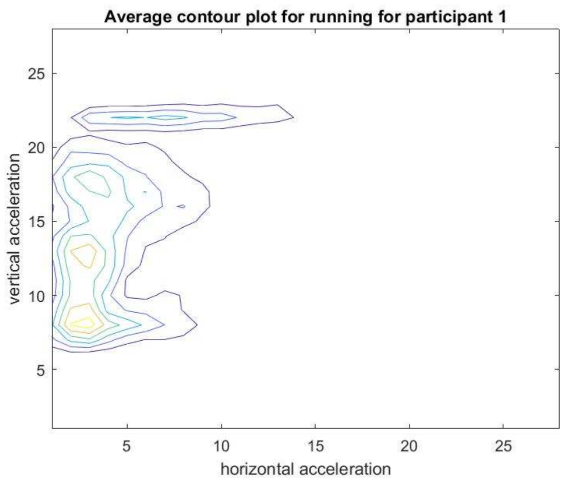

Fig. 3. Average contour plot for the running acceleration images, participant 1

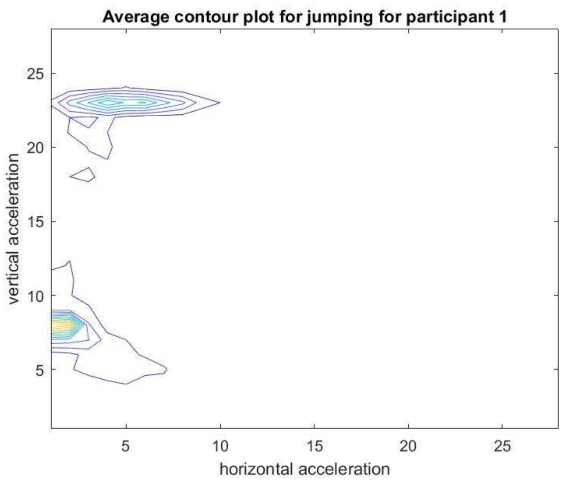

Fig. 4. Average contour plot for the jumping acceleration images, participant 1

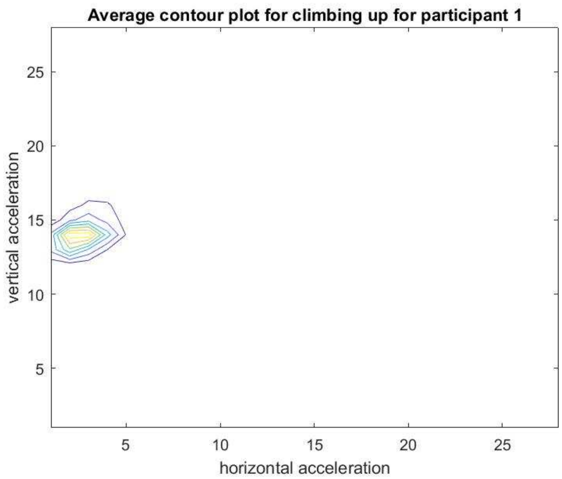

Fig. 5. Average contour plot for the climbing up acceleration images, participant 1

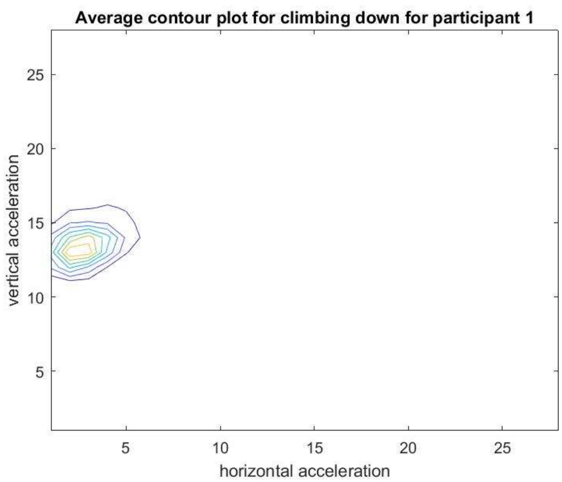

Fig. 6. Average contour plot for the climbing down acceleration images, participant 1

As previously stated, there are significant differences among the participants in the dataset in parameters such as their physical fitness, age, weight, height, and movement behavior [11]. These differences will result in different acceleration patterns and therefore in different acceleration images. Figures 7 to 11 capture the average monochrome HV acceleration images for participant 3 when executing the different activities. Although there are some similar characteristics with those in figures 2 to 6 there are also significant differences. The major impact of such a fact is that a L1O validation approach will provide worse results for those participants executing the activities in a different way from the rest of the participants. Training the CNN in section III with the acceleration images from participant 1 alone, for instance, and validating the CNN with acceleration images for participant 3 will provide worse results that using a p-fold cross-validation using data from both users in the training and validation phases.

Participant 1 in the dataset corresponds a 52-year-old female participant weighing 48 kg. Participant 3, on the other hand, corresponds to a 27-year-old male weighing 81 kg. The acceleration images from participant 3 show information at higher acceleration values for all the activities as compared to those from participant 1. The jumping activity for participant 3 also shows information on 2 distinct areas of the image corresponding to high positive and negative vertical accelerations. The running activity shows a wider connected region with content both for high vertical and horizontal accelerations. The climbing up activity also shows more prominent content for vertical positive accelerations than for the walking and climbing down activities.

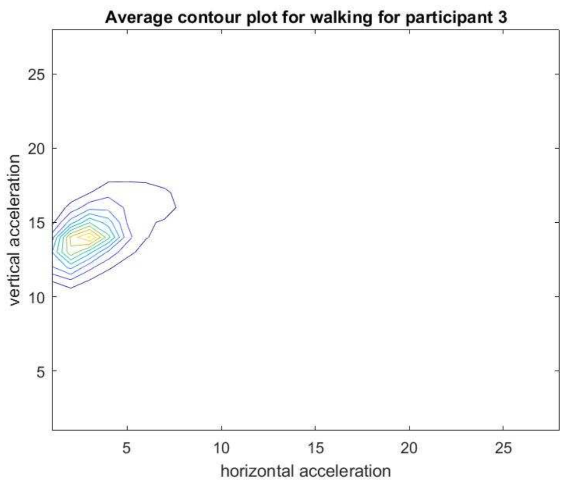

Fig. 7. Average contour plot for the walking acceleration images, participant 3

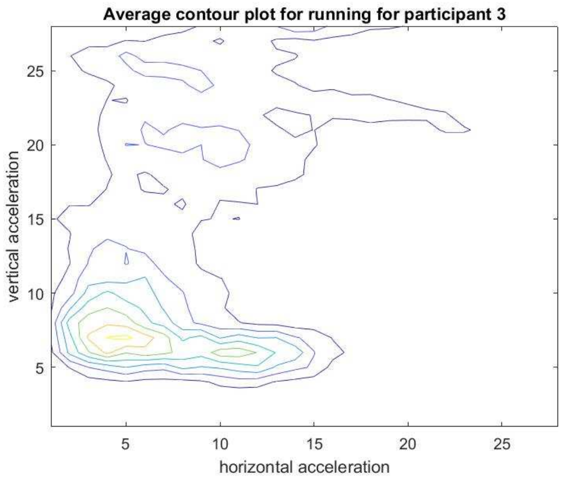

Fig. 8. Average contour plot for the running acceleration images, participant 3

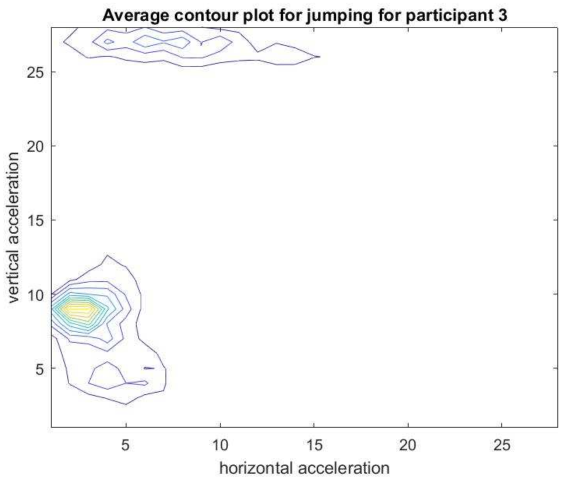

Fig. 9. Average contour plot for the jumping acceleration images, participant 3

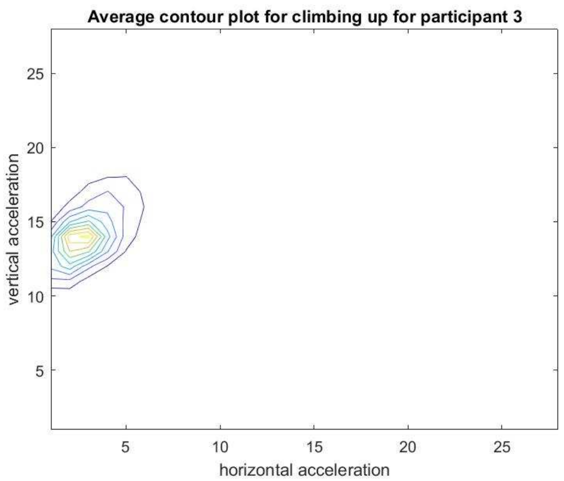

Fig. 10. Average contour plot for the climbing up acceleration images, participant 3

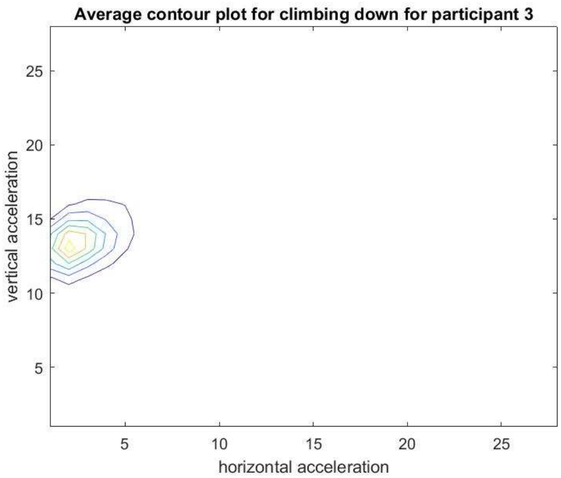

Fig. 11. Average contour plot for the climbing down acceleration images, participant 3

The following sub-sections will capture the results based on the type of HV image (monochrome or colored) and the validation mechanism (5-fold cross-validation or L1O) used in each case. In order to simplify the presentation of the information in the result tables, Table III assigns a symbol to each of the activities.

TABLE III
ACTIVITY SYMBOLS

Activity	Symbol
Walking	A1
Running	A2
Jumping	A3
Climbing up	A4
Climbing down	A5

In order to compare results with those presented in [11] using the same dataset, the F_1 score will be calculated as:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

A. Monochrome HV acceleration images and 5-fold cross-validation

The CNN described in section III was first trained with 80% of the information for all participants and validated with the remaining 20% (the process was repeated 5 times so that each acceleration window was used once and only once in the validation set). The tri-axial acceleration from a single sensor placed on the waist of each participant, sampled at 50 *Hz* and divided into 5 second windows was converted into 2D (Horizontal-Vertical) acceleration images as described in section III. The results for the confusion matrix, recall, accuracy and F_1 score values are captured in Table IV. The average accuracy is 0.94. The activities showing a best F_1 score value are the running and jumping activities (F_1=0.99). This result is aligned with the visual differences in the acceleration images from the jumping and running activities among themselves and with the rest of activities as shown in figures 2 to 11 and also aligned with the results in [11]. The worst F_1 score value was obtained for the walking/climbing up activity (F_1=0.88). There are some climbing up acceleration segments which are considered to be walking segments which makes sense since the ascending trajectory for some participants in the dataset included quasi-flat regions connecting ascending segments. The 2 activities with the worst precision figures (precision = 0.91) are in fact the walking and the climbing up since some of the acceleration segments in the climbing up recordings are detected as walking on a quasi-flat segment. The average F_1 score value is F_1=0.94.

TABLE IV
CONFUSION MATRIX MONOCHROME IMAGES 5-FOLD CROSS-VALIDATION

Activity	A1	A2	A3	A4	A5	Recall	F_1
A1	1256	0	0	54	5	0.96	0.93
A2	2	1473	0	9	3	0.99	0.99
A3	0	2	158	0	0	0.99	0.99
A4	107	11	0	978	53	0.85	0.88
A5	11	0	0	32	666	0.94	0.93
Precision	0.91	0.99	1.00	0.91	0.92		

Comparing the results in Tables I and IV, we observe a significant improvement in the F_1 score values using square HV acceleration images and a CNN as compared to the best classifier in [11] over the same dataset and using the major hand-crafted features previously used in related studies. In both cases, the information of a single accelerometer (in a known body location) and a p-fold cross-validation technique are used. The 2D (monochrome) acceleration images capture enough information to describe the human activities and the CNN automatically learns optimal features to recognize them. There is an 8% improvement on the average F_1 score values comparing tables I and IV and all the activities show better precision and recall values. The results in [11] also show that there are some climbing up segments classified as walking. Moreover, some climbing up segments are classified as climbing down and some walking segments are recognized as either climbing up or down activities. Using 2D square HV acceleration images to feed the CNN proposed in section III provides fewer walking segments miss-classified as climbing up or down.

B. Colored HV acceleration images and 5-fold cross-validation

Using the same validation procedure as in the previous sub-section but generating 3D (colored) images as described in section III provides the results in table V. Compared to the results in Table IV, there are small differences. Using colored images improve the recall for climbing up segments (fewer climbing up segments are classified as walking segments). However, more acceleration windows performed during a jumping activity execution are classified as running segments. The average F_1 score value is similar as the one obtained for monochrome images (F_1=0.94). One reason for this could be the fact that all 5 considered activities are periodic and their periods are not exact divisors of the time length of the 5-second window used to generate the acceleration images. There will be therefore 3D (colored) images covering the whole time-shifted versions of a randomly chosen acceleration window and the color will not provide additional information in this particular scenario.

TABLE V
CONFUSION MATRIX COLOR IMAGES 5-FOLD CROSS-VALIDATION

Activity	A1	A2	A3	A4	A5	Recall	F_1
A1	1251	2	0	56	7	0.95	0.93
A2	0	1482	2	2	2	1.00	0.99
A3	1	10	141	7	0	0.89	0.93
A4	106	2	0	999	42	0.87	0.88
A5	12	4	0	48	645	0.91	0.92
Precision	0.91	0.99	0.99	0.90	0.93		

Comparing tables I and V, a 7.7% average improvement in the F_1 score value is obtained when using the 3D (colored) HV acceleration images and a CNN to learn the best describing features as compared to the best classifier and hand-crafted features used in [11] with the same dataset.

C. Monochrome HV acceleration images and L1O validation

As captured in the introduction section of this paper, and according to previous research studies, a leave on out (L1O) approach will provide worse results than a p-fold cross-validation multi-user schema. The difference will be small when similar users exist in the dataset and big in the case of users executing the activities in a peculiar way not used by the rest of the users. The authors in [11] proposed classifying the users into similar clusters in order to validate each user with a learning algorithm trained with data from similar users. In fact, the data in the dataset used [27] includes an important heterogeneity both in terms of users (age, height, weight, fitness level) and the environment in which the activities are performed (running on a treadmill, in the city or in the countryside for example). This sub-section captures the results when using a L1O approach over 2D (monochrome) HV acceleration images. The results when using colored images will be presented in the next sub-section.

Table VI captures the recall, precision and F_1 score values for the 5 activities using 2D HV acceleration images and a L1O validation approach. In this case, the number of walking segments classified as climbing up increases significantly. Some running segments are considered as executed in a jumping activity. The number of climbing up segments classified as walking also increase. The best recognition in terms of the computed F_1 score value is achieved for the running activity. The overall accuracy is in this case 0.77.

TABLE VI
CONFUSION MATRIX MONOCHROME IMAGES L1O VALIDATION

Activity	A1	A2	A3	A4	A5	Recall	F_1
A1	783	6	1	427	98	0.60	0.67
A2	0	1330	136	1	20	0.89	0.92
A3	0	11	149	0	0	0.93	0.67
A4	225	26	0	836	62	0.73	0.67
A5	20	17	0	72	600	0.85	0.81
Precision	0.76	0.96	0.52	0.63	0.77		

Comparing the values with those in table II (as described in [11] for the data from the same accelerometer in the waist of the participants) there is only a very small improvement in the F_1

score value of 0.13% when using the CNN trained with 2D acceleration images as compared to the best classifier and hand-crafted features in [11]. The intuition is that the CNN will only be able to learn the best classification features based on the information in the training set. For particular users (statistically different from those in the training set) the CNN will not be able to provide accurate results (as any other learning method).

D. Colored HV acceleration images and L1O validation

This last sub-section will present the results achieved when using a L1O validation approach over 3D (colored) square HV acceleration images. The results are captured in table VII. In this case, the F_1 score values slightly improve as compared to those when using 2D (monochrome) images. There are less walking acceleration segments which get misclassified as climbing down sections. There are fewer running windows miaclassified as jumping. There is however a small degradation in the jumping recognition. There is also a slight improvement in the recognition of climbing up and climbing down segments. A smaller number of climbing up acceleration windows are recognized as waking segments and climbing down sections are less confused with running segments.

TABLE VII
CONFUSION MATRIX COLOR IMAGES L1O VALIDATION

Activity	A1	A2	A3	A4	A5	Recall	F_1
A1	815	4	0	444	52	0.62	0.69
A2	0	1348	110	2	27	0.91	0.94
A3	2	12	133	6	6	0.84	0.66
A4	189	22	1	873	65	0.76	0.68
A5	27	4	0	82	596	0.84	0.82
Precision	0.79	0.97	0.55	0.62	0.80		

Comparing the results in table VII with those in table II, there is an improvement of 1.6% in the average F_1 score value.

VI. CONCLUSION

Deep Learning approaches to Human Activity Recognition (HAR) are providing better results than those based on hand-crafted features. This paper proposes a novel mechanism to transform the data obtained from a single tri-axial accelerometer into square acceleration images suitable to train a Convolutional Neural Network (CNN) for HAR. Both a monochrome and a colored version are provided. Monochrome images capture the vertical and horizontal (geo-referenced) accelerations in a window of data. The time information is added as color in the 3D (colored) images.

A multi-user validation approach using both a p-fold cross-validation and L1O validation schema are used. A publicly available dataset comprising a heterogeneous set of users performing a set of activities in different environments has been used to validate the generalization of results. The activities to be recognized are climbing stairs down and up, jumping, running/jogging and walking. A single tri-axial accelerometer located in the waist region of the participants has been used.

The results for the p-fold cross-validation approach outperforms the best previous results using the same dataset by around 8% (as for the F_1 score values). There are no significant differences between the monochrome and colored approaches mainly due to the periodic characteristics of the activities to be recognized. From the 5 activities to be classified, the jumping and running activities showed visually differentiated characteristics when examining the generated acceleration images as opposed to the other 3 activities. The results confirm that the F_1 score values achieved for the jumping and running activities outperformed those for the other 3 activities.

The results achieved using square HV acceleration images for the L1O validation approach only slightly improve previous results. In this case, the colored version outperforms the monochrome version by around 1% in terms of the F_1 score values. Leaving out a participant and training the CNN with the information from the rest of the participants in the training set will only provide accurate results if there are similar participants in the training set as the one left out. The particularities of the selected dataset included individuals performing the activities with certain particularities that makes it difficult for any machine learning algorithm to guess.

As future work, the proposed algorithms will be applied to other public datasets including a richer set of activities and some variations to the parameters of the algorithms such as the size of the acceleration images will be explored. Moreover, the algorithms proposed in this paper will be applied to the same dataset in [27] but to the other sensor locations to compare the results with those presented in this paper for the waist sensor.

ACKNOWLEDGMENT

This work was supported by the "ANALYTICS USING SENSOR DATA FOR FLATCITY" project TIN2016-77158-C4-1-R (Ministerio de Ciencia, innovación y Universidades/ ERDF, EU) funded by the Spanish Agencia Estatal de Investigación (AEI) and the European Regional Development Fund (ERDF)."

REFERENCES

[1] L. Chen, J. Hoey, C.D. Nugent, D.J. Cook, Z. Yu, "Sensor-based activity recognition" IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., 42 (6) (2012), pp. 790-808

[2] O.D. Lara, M.A. Labrador, "A survey on human activity recognition using wearable sensors", IEEE Commun. Surv. Tutor., 15 (3) (2013), pp. 1192-1209

[3] A. Bulling, U. Blanke, B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors", ACM Comput. Surv., 46 (3) (2014), p. 33

[4] N. Ravi, N. Dandekar, P. Mysore, M.L. Littman, "Activity recognition from accelerometer data", Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence - Vol. 3, IAAI'05, AAAI Press (2005), pp. 1541-1546

[5] L. Bao, S.S. Intille, "Activity recognition from user-annotated acceleration data", Pervasive Computing, Springer (2004), pp. 1-17

[6] Jiang, Wenchao, and Zhaozheng Yin. "Human activity recognition using wearable sensors by deep convolutional neural networks." Proceedings of the 23rd ACM international conference on Multimedia. ACM, 2015.

[7] G.M. Weiss, J.W. Lockhart, The impact of personalization on smartphone-based activity recognition, in: AAAI Workshop on Activity Context Representation: Techniques and Languages, 2012, pp. 98–104.

[8] K. Altun, B. Barshan, "Human activity recognition using inertial/magnetic sensor units", in International Workshop on

Human Behavior Understanding, Springer, Berlin, Heidelberg (2010), pp. 38-51
[9] Z. Wang, M. Jiang, Y. Hu, H. Li, "An incremental learning method based on probabilistic neural networks and adjustable fuzzy clustering for human activity recognition by using wearable sensors" in IEEE Trans. Inf. Technol. Biomed., 16 (4) (2012), pp. 691-699
[10] D. Coskun, O.D. Incel, A. Ozgovde, "Phone position/placement detection using accelerometer: Impact on activity recognition", in 2015 IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing, (ISSNIP), IEEE (2015), pp. 1-6
[11] Sztyler, T., Stuckenschmidt, H., & Petrich, W. (2017). Position-aware activity recognition with wearable devices. Pervasive and mobile computing, 38, 281-295.
[12] Jordao, A., Torres, L. A. B., & Schwartz, W. R. (2018). Novel approaches to human activity recognition based on accelerometer data. Signal, Image and Video Processing, 1-8. https://link.springer.com/article/10.1007/s11760-018-1293-x
[13] Ha, S., Choi, S.: Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In: IJCNN (2016)
[14] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).
[15] Bruno, B., Mastrogiovanni, F., Sgorbissa, A.: Wearable inertial sensors: applications, challenges, and public test benches. IEEE Robot. Autom. Mag. 22, 116–124 (2015)
[16] O. Banos, R Garcia, J, A. Holgado-Terriza, M Damas, H Pomares, I Rojas, A Saez, C. Villalonga, "mHealthDroid: A novel framework for agile development of mobile health applications", Proceedings of the International Work-Conference on Ambient Assisted Living (IWAAL), 2014.
[17] Ha, S., Yun, J., Choi, S.: Multi-modal convolutional neural networks for activity recognition. In: IEEE SMC (2015)
[18] O. Banos, R Garcia, J, A. Holgado-Terriza, M Damas, H Pomares, I Rojas, A Saez, C. Villalonga, "mHealthDroid: A novel framework for agile development of mobile health applications", Proceedings of the International Work-Conference on Ambient Assisted Living (IWAAL), 2014.
[19] P. Zappi, C. Lombriser, T. Stiefmeier, E. Farella, D. Roggen, L. Benini, and G. Troster, "Activity recognition from on-body sensors: accuracy- ¨ power trade-off by dynamic sensor selection," in Wireless sensor networks. Springer, 2008, pp. 17–33
[20] Jiang, W., Yin, Z.: Human activity recognition using wearable sensors by deep convolutional neural networks. In: ACM Conference on Multimedia Conference (2015)
[21] Davide Anguita, et al., "A public domain dataset for human activity recognition using smartphones," in European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN, 2013.
[22] Mi Zhang and Alexander A Sawchuk, "Usc-had: a daily activity dataset for ubiquitous activity recognition using wearable sensors," in ACM Conference on Ubiquitous Computing, 2012.
[23] Muhammad Shoaib, et al., "Fusion of smartphone motion sensors for physical activity recognition," Sensors, 2014.
[24] Chen, Y., Xue, Y.: A deep learning approach to human activity recognition based on single accelerometer. In: IEEE SMC (2015)
[25] Mathworks, "Matlab". Last visited on July 2018. Available on-line: https://es.mathworks.com/products/matlab.html
[26] Yann LeCun, Corinna Cortes, Christopher J.C. Burges, The MNIST Database of handwritten digits. Available on-line: http://yann.lecun.com/exdb/mnist/. Last visited on 27 June 2018
[27] Timo Sztyler, Heiner Stuckenschmidt and Wolfgang Petrich, "On-body Localization of Wearable Devices: An Investigation of Position-Aware Activity Recognition". Last visited on July 2018. Available on-line: https://sensor.informatik.uni-mannheim.de/#dataset_realworld
[28] D. Mizell, Using gravity to estimate accelerometer orientation, in: Proceedings of the Seventh IEEE International Symposium on Wearable Computers, 2003, 2003, pp. 252–253.
[29] Pham, C. (2015, October). MobiRAR: real-time human activity recognition using mobile devices. In Knowledge and Systems Engineering (KSE), 2015 Seventh International Conference on (pp. 144-149). IEEE.

M. Muñoz Organero (M'08) received the M.Sc. degree in telecommunications engineering from the Polytechnic University of Catalonia, Barcelona, Spain, in 1996, and the Ph.D. degree in telecommunications engineering from the Carlos III University of Madrid, Leganes, Spain, in 2004.
He is a professor of telematics engineering at the Carlos III University of Madrid where he is currently the head of the Telematics Engineering Departement. From September 2015 to August 2016 he was a visiting professor in the CATCH Centre in the University of Sheffield. His research projects have included topics related to ambient intelligence, human activity recognition, machine learning techniques applied to human behavior, health and independent living systems, ITS, open architectures for e-learning systems, open service creation environments for next-generation networks, advanced mobile communication systems, pervasive computing, and convergent networks. He has published more than 160 research papers both on journals and international conferences. He has participated in several European-funded projects such as E-LANE and Spanish-funded projects such as MOSAIC learning, Learn3 and OSAMI. He is currently the PI of the Spanish-funded FLATCity project and was PI on the HERMES, ARTEMISA, HAUS, COMINN, REMEDISS and IRENE projects. He was also PI for the GEEWHEZ EU-FP7 project and an MC member of TU1305 EU funded COST action about social networks and travel behaviour.