CCF Transactions on Pervasive Computing and Interaction (2020) 2:18–32
https://doi.org/10.1007/s42486-020-00026-2

REGULAR PAPER

Feature learning for Human Activity Recognition using Convolutional Neural Networks

A case study for Inertial Measurement Unit and audio data

Federico Cruciani[1] · Anastasios Vafeiadis[2] · Chris Nugent[1] · Ian Cleland[1] · Paul McCullagh[1] · Konstantinos Votis[2] · Dimitrios Giakoumis[2] · Dimitrios Tzovaras[2] · Liming Chen[1] · Raouf Hamzaoui[3]

Received: 11 November 2019 / Accepted: 4 January 2020 / Published online: 24 January 2020

Abstract

The use of Convolutional Neural Networks (CNNs) as a feature learning method for Human Activity Recognition (HAR) is becoming more and more common. Unlike conventional machine learning methods, which require domain-specific expertise, CNNs can extract features automatically. On the other hand, CNNs require a training phase, making them prone to the cold-start problem. In this work, a case study is presented where the use of a pre-trained CNN feature extractor is evaluated under realistic conditions. The case study consists of two main steps: (1) different topologies and parameters are assessed to identify the best candidate models for HAR, thus obtaining a pre-trained CNN model. The pre-trained model (2) is then employed as feature extractor evaluating its use with a large scale real-world dataset. Two CNN applications were considered: Inertial Measurement Unit (IMU) and audio based HAR. For the IMU data, balanced accuracy was 91.98% on the UCI-HAR dataset, and 67.51% on the real-world Extrasensory dataset. For the audio data, the balanced accuracy was 92.30% on the DCASE 2017 dataset, and 35.24% on the Extrasensory dataset.

Keywords Convolutional Neural Networks · Deep learning · Human Activity Recognition · Free-living

1 Introduction

In recent years, research in Machine Learning (ML) has gone through some of its biggest advancements. In particular, Deep Learning (DL) methods have brought significant

This research has received funding under the ACROSSING project Marie Skłodowska-Curie EU Framework for Research and Innovation Horizon 2020, under Grant Agreement No. 676157.

✉ Federico Cruciani
f.cruciani@ulster.ac.uk

Anastasios Vafeiadis
anasvaf@iti.gr

Chris Nugent
cd.nugent@ulster.ac.uk

Ian Cleland
i.cleland@ulster.ac.uk

Paul McCullagh
pj.mccullagh@ulster.ac.uk

Konstantinos Votis
kvotis@iti.gr

Dimitrios Giakoumis
dgiakoum@iti.gr

Dimitrios Tzovaras
tzovaras@iti.gr

Liming Chen
l.chen@ulster.ac.uk

Raouf Hamzaoui
rhamzaoui@dmu.ac.uk

[1] School of Computing, Ulster University, Newtownabbey, UK

[2] Information Technologies Institute, Center of Research and Technology Hellas, Thessaloniki, Greece

[3] Faculty of Computing, Engineering and Media, De Montfort University, Leicester, UK

improvements in several fields where ML models are currently employed. The first applications of DL methods have been in computer vision and natural language processing (LeCun et al. 2015). The accuracy improvement brought by such methods has caused an increasing popularity; nonetheless, their application to Human Activity Recognition (HAR) is relatively new. Consequently, the exploration of use of DL in HAR provides scope for significant contribution. There is no general definition of DL, making it difficult to classify HAR methods in this perspective. In LeCun et al. (2015), the authors highlighted one of the main characteristics distinguishing DL methods with respect to conventional ML: i.e., the capacity of Convolutional Neural Network (CNNs) of learning data representation in an automatic fashion.

Studies on HAR have been undertaken over the last two decades (Bulling et al. 2014). Conventional ML and the use of Human Crafted Features (HCF) for HAR have been deeply investigated, for instance in Janidarmian et al. (2017); Espinilla et al. (2018). Those studies evaluated feature selection strategies for HAR, and helped to identify which features are more relevant depending on the set of target activities, type of sensors, and sensor's location where relevant. Recent studies have shown that CNN-based automatic feature extraction can provide results comparable to the best known HCF case (Ronao and Cho 2016). Nevertheless, most studies have focused on comparing the HCF and the CNN case based on final accuracy of a trained classifier, for instance (Li et al. 2018). Moreover, in most cases, HAR methods have been evaluated using data collected in controlled environments, i.e., with data that are possibly under-representing the main challenges that real-world deployment introduces (Vaizman et al. 2017). In contrast, this work aims at evaluating CNNs as a feature extractor in a real-world environment. This article is an extended version of the work published in Cruciani et al. (2019b). In our previous work, we compared the performances of using HCF and CNN automatic features, and explored the effect of the main hyperparameters on the feature learning abilities of CNNs. This work aims at providing a real-world evaluation of CNN as feature extractors for HAR, considering two different sensor modalities: Inertial Measurement Unit (IMU)-based and audio based. This paper makes the following contributions:

1. An evaluation of the best identified CNN architecture for IMU and audio based HAR is performed using data collected in controlled conditions.
2. An evaluation of the identified CNN architecture is performed on a large real-world publicly available dataset.

The remainder of this paper is structured as follows. Section 2 provides an overview of related work, highlighting the contribution of this study in the context of past studies. Section 3 describes the proposed case study employing CNN-based feature extractors for HAR. Section 4 describes the experiments undertaken and the evaluation methodology. Results and discussion are reported in Sects. 5 and 6 respectively. Finally, conclusions are drawn in Sect. 7.

2 Related work

The generic Activity Recognition Chain (ARC) (Bulling et al. 2014) for conventional supervised ML approaches, as depicted in Fig. 1, consists of four steps leading from raw data to activity classification; namely: pre-processing, segmentation, feature extraction, and classification. Some DL methods, as in the case of CNNs, allow classification directly with pre-processed segmented raw-data, without requiring an explicit step for feature extraction (LeCun et al. 2015). Feature extraction, in the case of CNNs, is performed through the convolution of the input signal with a kernel (also referred to as filter) (Ordóñez and Roggen 2016). The result of the convolution operation is known as feature map (Ordóñez and Roggen 2016). The ability of CNNs to learn features automatically has a twofold consequence. On the one hand, it simplifies the ARC by automating a step that typically requires significant domain-specific expertise to identify a suitable feature set (LeCun et al. 2015). This process is usually accomplished by applying a feature selection strategy: starting from the largest possible set of features, to then reduce the feature set to the ones providing better discrimination between target classes; whereas for CNNs none of these steps are required. On the other hand, the use of CNNs moves the feature extraction step to within the classifier model, meaning that a CNNs feature extractor requires a training phase in order to generate suitable features, exposing the approach to the cold-start problem. This drawback of using CNNs has often been addressed in computer vision, where it is common to use pre-trained CNN models for feature extraction, for instance in Rajaraman et al. (2018).

Figure 2 illustrates the difference between the HCF case and the use of CNN for feature extraction in the case of a Multi-Layer Perceptron (MLP) classifier. The MLP classifier consists of a series of dense fully connected layers,

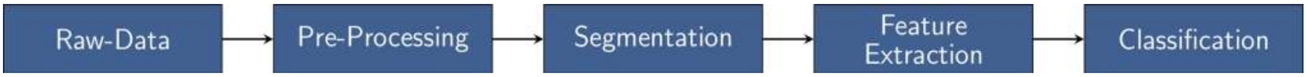


Fig. 1 The Activity Recognition Chain (ARC) in conventional ML approaches. Adapted from Bulling et al. (2014)

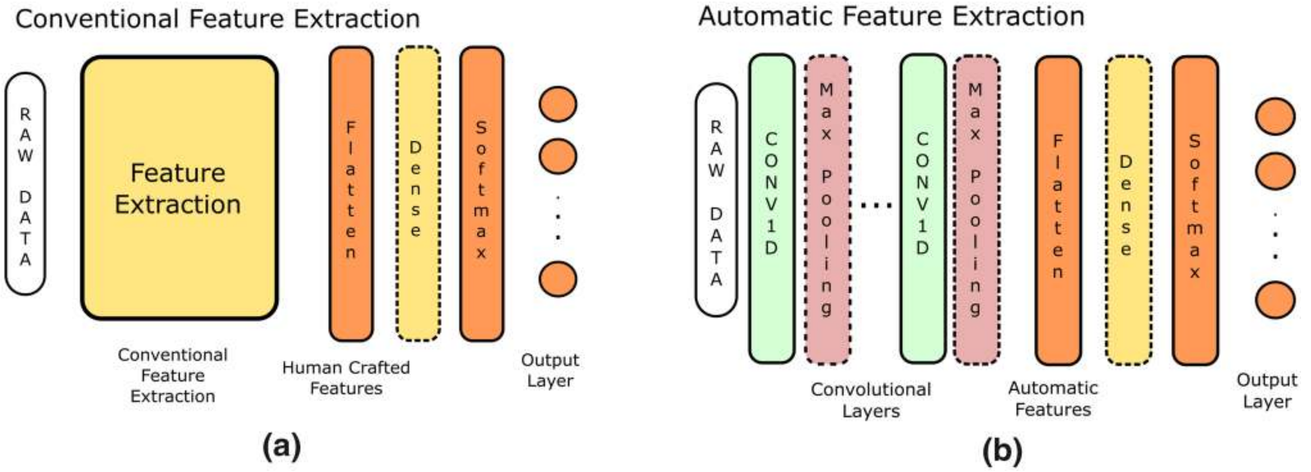


Fig. 2 Feature extraction step in the case of HCF and with CNN automatic features. Adapted from Cruciani et al. (2019b)

leading to an output layer with the same number of nodes as the number of target classes. In the conventional case (a), the MLP is fed with an input consisting of a vector of HCF. In the CNN case (b), a series of convolutional layers accomplish the step of feature extraction (Baldominos et al. 2019).

The convolutional layers composing the architecture can be optionally followed by a max-pooling operation, with the goal of down-sampling the data representation, thus reducing the size of the feature map (Baldominos et al. 2019). After the series of convolutional layers, the output of last convolutional layer is usually flattened into a 1D vector, feeding the MLP as in the HCF case. Convolutional layers are typically implemented as 2D convolution, as in image (and sometimes) audio processing; whereas 1D (or temporal) convolution is more common for IMU signals (Moya Rueda et al. 2018; Saeed et al. 2018). Figure 3 visualizes an example of temporal convolution where the kernel size is 2.

Rectified Linear Unit (ReLU) is among the most common activation function for convolutional layers, whereas, for connecting the last dense layer with the output layer, Softmax is typically used in multi-class classification problems (i.e., when output classes are mutually exclusive) (Ordóñez and Roggen 2016). Other activation functions such as sigmoid can be used in the case of multi-label classification (i.e., when more than one output node can be active at the same time e.g., 'sitting' and 'on a bus') (Huang et al. 2019).

In this work, two main applications of CNNs are considered, both with application to HAR, namely: IMU-based, and audio-based HAR. To shorten the notation of the CNN architectures, we use *n*-CNN where *n* is the number of convolutional layers, with *k* indicating the size of the kernel, and *f* denoting the number of filters (or kernels).

2.1 IMU-based methods

IMU are among the most investigated sensor modalities to perform HAR (Bulling et al. 2014). In contrast to vision-based systems, for instance, inertial sensors do not pose privacy issues, are available on-board all modern smartphones, and are more energy efficient than other sensors such as the GPS.

HCFs for IMU-based methods have been deeply investigated in the past; with studies that identified the most relevant features for a range of specific settings. In particular, these studies investigated the best HCF sets, depending on the set of target activities, and sensor location (Morales and Akopian 2017; Janidarmian et al. 2017; Espinilla et al. 2018). With DL becoming more popular, recent studies have started to analyze the case of DL in comparison with conventional ML (Li et al. 2018; Baldominos et al. 2019). In these cases, the comparison mostly focused on the evaluation of the final accuracy of models, using different feature learning strategies, including HCF and CNNs. In Ronao and Cho (2016), a more detailed analysis of CNNs was provided for IMU-based HAR. The authors analyzed the impact of the main hyperparameters such as, the number of convolutional layers, and the kernel size used for the convolution. Compared to Ronao and Cho (2016) and Cruciani et al. (2019b),

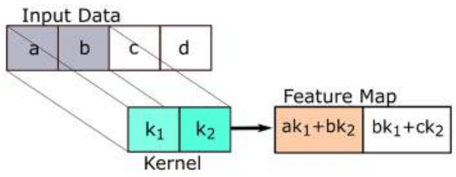


Fig. 3 Feature map obtained using 1D convolution and kernel size of 2. Adapted from Baldominos et al. (2019)

the current work has evaluated CNNs on a more challenging set of target activities (including a NULL class), and evaluated a pre-trained feature extractor in realistic conditions.

2.2 Audio-based methods

Audio-based event recognition has received significant research attention in the last ten years and many public datasets have been released (Gemmeke et al. 2017; Mesaros et al. 2017) to help researchers benchmark their algorithms. The low cost of microphone sensors and the high processing power of single-board computers has increased the interest for on-device processing for various applications, specifically for real-time remote tracking of patients from their health care providers (Alsina-Pagès et al. 2017).

One of the fundamental problems in the audio-based event recognition is the feature extraction. Many types of low-level features such as zero-crossing rate, band-energy ratio, spectral roll-off, spectral flux, spectral centroid, spectral contrast, mel-frequency cepstral coefficients (MFCCs) and gammatone frequency cepstral coefficients are commonly used in the literature (Peltonen et al. 2002; Eronen et al. 2006; Perttunen et al. 2008; Valero and Alias 2012; Xia et al. 2018). Most of the aforementioned features work well for specific datasets but may fail on others. For instance, the MFCCs provide good classification results for the speech recognition task, but can have poor performance when classifying unstructured, noisy data, such as environmental sounds. The main reason is that the MFCCs convert the input signal to the mel-scale, using a log operation on the power spectrum, which relates to how the human ear perceives sounds (Zhao and Wang 2013). Therefore, there could be frequencies that are not emphasized, which are important for environmental sound.

Recently, deep CNNs have been successful in many tasks, such as speech recognition (Abdel-Hamid et al. 2014), audio source separation (Grais et al. 2018), environmental sound recognition (Morfi and Stowell 2018) and end-to-end polyphonic event detection (Çakir and Virtanen 2018). The fundamental difficulty of environmental sound recognition is that the input signal is highly variable due to different environments (indoor, outdoor, vehicle) and acoustic conditions (echo, reverb).

2.3 Contribution of this work

With respect to past studies using CNNs, this work presents the following contributions. Firstly, two example cases, IMU-based and audio-based methods are considered. Although, other studies already compared different feature learning methods including CNNs, the comparison is typically limited to the assessment of final accuracy performance in controlled conditions. In this study the best candidate CNN architectures for audio and IMU-based HAR are evaluated under realistic conditions, using a large real-world public dataset.

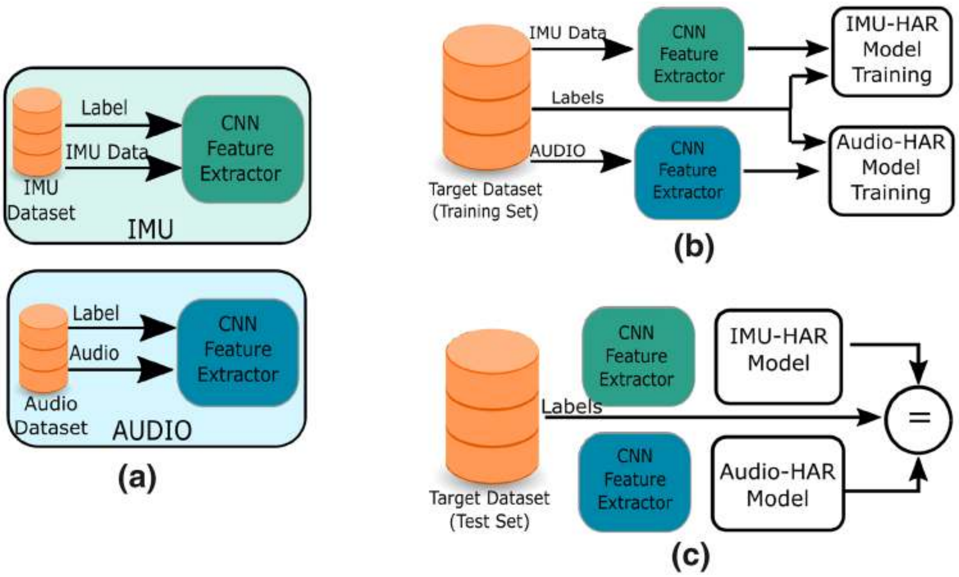
3 Case study

The current work aimed at experimenting the use of CNN with a twofold goal. The first goal of this study was to analyze the quality of CNN automatically extracted features, with different hyperparameters and topologies. The second goal was to explore the use of a pre-trained CNN feature extractor on a real-world dataset for HAR. Such an approach benefits from the CNN ability to automate feature extraction, while at the same time avoiding the cold-start problem.

In our previous work (Cruciani et al. 2019b), a case study was proposed composed of two steps, as depicted in Fig. 4. In the first step (Fig. 4a), a CNN feature extractor is trained. In this step, the effect of different topologies and hyperparameters combinations is evaluated. This analysis identifies the best performing models for HAR. In the second step, the best performing CNN model, trained in the previous step, is used only as feature extractor, converting raw-data into a suitable input vector for a second classifier model. Weights of the CNN networks are frozen in the first step, and the CNN model is used by taking the feature vector produced by the flatten layer after the series of convolutional operations. By taking the output of the flatten layer, the feature vector obtained can be used as a representation of raw data in a different context, following a paradigm similar to transfer learning. Features generated by this pre-trained classifier are used to train the second model, as presented in Fig. 4b. Finally, performance of the second model is evaluated, as presented in Fig. 4c.

In Cruciani et al. (2019b), some preliminary experiments were conducted to compare HCF and CNN; however, the scope was limited to the first step of the current case study. In this work, we report on the completion of the second step in form of an illustrative example on how to use a CNN feature extractor for HAR, together with some additional results and analysis regarding the first step, not present in the previous work. In the previous work, some basic requirements regarding suitable datasets for the case study were identified. In particular, it was identified that for the first step the use of datasets collected in controlled environments was preferable. The nature of such datasets, being collected in controlled conditions, does not expose to the risk of noisy labels, that may occur when the annotation occurs in uncontrolled conditions. Issues like label noise may affect the comparison with HCFs leading to an incorrect evaluation. Consequently, two datasets were identified: the UCI-HAR (Anguita et al. 2013) for the IMU, and the DCASE 2017 (Mesaros et al. 2017) for the audio, as controlled environment datasets to

Fig. 4 The proposed case study: **a** the CNN-based feature extractor is trained on a dataset and obtained model weights are frozen, **b** automatic features extracted using the pre-trained CNN feature extractor are used to train a second classifier on a different dataset; finally **c**, testing of the second model allows the evaluation of the final classification accuracy on the new dataset. Adapted from Cruciani et al. (2019b)



be used in the first step. The second step of our case study aimed at evaluating the use of the pre-trained CNN feature extractor, this time, in real-world conditions. Naturally, the datasets used in step 1 and step 2 must have similar characteristics: with input data of the same nature, and similar target activities (although the set of target activities may differ between the two cases).

In implementing our case study for the IMU sensor, a classifier of different nature (i.e., not CNN based) was chosen as second model for the IMU. A Random Forest (RF) model was identified as a suitable model, using CNN extracted features instead of HCF. RF were proven to be among the best classification methods for IMU-based HAR (Baldominos et al. 2019), and are commonly used for HAR. Based on these findings an RF model was used, although any supervised classifier could be used in this step, and the choice of RF simply aims at providing an illustrative example. Regarding the audio modality, a 1D CNN was used for our case study. For both the IMU and audio case, evaluation of the models was performed using the Exatrasensory dataset (Vaizman et al. 2017), providing the challenging case of a real-world dataset.

4 Experiment

The experiment was conducted following the two steps proposed in the case study:

1. Comparing the performance of different CNN architectures and hyperparameters. As result of the comparison, two CNN feature extractors are trained: one for the IMU and one for the audio case.
2. A second model is trained using features extracted with the CNN pre-trained in step 1. Finally, evaluation of the second model using CNN features is performed in realistic conditions.

The following sections provide details of the experimental methods for each of these steps.

4.1 Step 1: Comparing CNN architectures

Using the two datasets identified for the IMU and audio cases, a number of CNN structures and hyperparameters were evaluated, in particular:

1. The number of convolutional layers
2. The kernel size used for the convolution
3. The number of filters

4.1.1 Evaluation methodology

The evaluation of hyperparameters and CNN architectures was performed considering that the more complex the model is, the higher are the chances of overfitting, particularly when training with small size datasets, or with data not universally representative of the target activities. Therefore, the evaluation of different models corresponding to different combinations of the aforementioned hyperparameters was undertaken starting from a simple model; and then gradually increasing the complexity in search of an optimal trade-off between accuracy and model complexity. The comparison evaluated the effect of the number of layers of convolution n, keeping the kernel size k fixed to $k = 2$ for the 1D convolution, and a relatively small number of filters ($f = 12$) in the first layer. The best combination was identified before proceeding to the next exploration phases: i.e., maintaining the number of convolutional layers n fixed, while increasing the kernel size $k = [2, 4, 8, 16, 32, 64]$, and, finally evaluating use of multiple filters $f = [12, 24, 48, 96, 128]$.

Evaluation using the accuracy measured on the final model was complemented by means of visualization of the features produced. For visualization purposes, Principal Component Analysis (PCA) was used to reduce the dimension of the feature space to the first three principal components, and generating two plots for each configuration showing the 1st and 2nd, and the 1st and 3rd components, respectively. As will be described in Sects. 5 and 6, the plot visualization helped with interpretation of the results during the exploration of the different combinations of hyperparameters.

The best identified models were then used as feature extractor for step 2 of the case study.

4.1.2 IMU case

As previously mentioned, the UCI-HAR dataset was used for the IMU case, performing a comparison similar to Ronao and Cho (2016). Compared to Ronao and Cho (2016) and our previous work (Cruciani et al. 2019b), in this work experiments were conducted using the updated version of the same dataset (Reyes-Ortiz et al. 2016), which includes labels also for postural transitions. Samples corresponding to transitions were used to define a more challenging target activity set, with transitions considered as the NULL class. This version of the dataset is publicly available.[1] The dataset includes tri-axial accelerometer and gyroscope signals recorded using a smartphone (Samsung Galaxy S2). The dataset also provides a set of 561 HCFs extracted from the accelerometer and gyroscope signals. This set of HCFs was used in our previous experiment comparing CNN and HCF features. The features available with the dataset were extracted using a window size of 128 samples (corresponding to 2.56 s with the 50 Hz sampling rate). The same segmentation was kept for this experiment. In this configuration the input layer of the CNN takes a 128 × 6 input shape (corresponding to the 3 channels X, Y, Z of the accelerometer and gyroscope signals). The dataset provides a train-test partition, with 21 of the 30 subjects as training set, and the remaining 9 subjects for testing purposes. To reduce the probability of overfitting, the 21 training subjects were divided into two groups 18 for training and 3 for validation during training. Evaluation of different configurations was made using Adam (Kingma and Ba 2015) and Stochastic Gradient Descent (SGD) optimizers, with different number of layers, kernel sizes, and number of filters. For final training of the CNN feature extractor, the SGD optimizer was used. Compared to the Adam optimizer, SGD provides, in some cases, better generalization on unseen data (Keskar and Socher 2017), and that is also for the case of IMU data for HAR (Cruciani et al. 2019a). The training stops when loss on the validation set stops decreasing. SGD typically causes more oscillations during the training process, thus, requiring a higher number of epochs to converge. Despite a slower training process, SGD provides two main advantages. Firstly, the stochastic approach increases the chances to improve over local minima solutions. Secondly, this reduces the risk of stopping the training process too early, ensuring that the model has gone through a higher number of epochs. The final training of the CNN feature extractor was performed on a high number of epochs ($\geq$ 10000) keeping a high patience (1000 epochs) and saving only best weights minimizing the loss on the validation set. On top of using a different optimizer, some additional variants were introduced in the final CNN model for feature extraction. These changes were made considering the different nature of the two datasets. In the UCI-HAR dataset, sensor location was constrained to the waist, whereas in the Extrasensory dataset it is uncontrolled. Consequently, two more channels (the 3D magnitude of the accelerometer and gyroscope) were added as input. This provided a 128 × 8 input to the CNN, introducing two rotation invariant channels. On top of that, accelerometer and gyroscope data were recorded at different sampling rates in the two datasets: 50 Hz for the UCI-HAR, and 40 Hz for the Extrasensory. In the final training, UCI-HAR data were down-sampled to 40 Hz in order to train the CNN feature extractor in a compatible manner with the Extrasensory data. At this stage, the identified optimal kernel size $k = 32$ (identified in the first step) was also adjusted to $k = 24$, in order to maintain the size of the filter in a comparable time length (about 0.5s) while switching from a 50 to 40 Hz sampling rate. This configuration was used to train the final version of the CNN feature extractor for the IMU on the UCI-HAR dataset as a 3-CNN with $k = 24$ and $f = 24$, where f is the number of filters used in the first layer. In the final structure for the second and third layers, 48 and 96 filters were used respectively; doubling the number of filters after each convolution, while using max-pooling to maintain the feature map to an equivalent size. The weights of the CNN were frozen at this stage. The CNN feature extractor obtained was used to process accelerometer and gyroscope data from the Extrasensory. Since the goal of our case study was to reuse the obtained pre-trained classifier, the use of dropout was introduced for model training at this stage, in an attempt of further reducing overfitting phenomena (Srivastava et al. 2014). A dropout layer (with rate set to 0.5) was added after each convolutional layer.

4.1.3 Audio case

The performance of the CNNs was evaluated on a large-scale dataset (Mesaros et al. 2017). The DCASE 2017

[1] http://archive.ics.uci.edu/ml/datasets/Smartphone-Based+Recognition+of+Human+Activities+and+Postural+Transitions.

development dataset consists of recordings from various acoustic scenes, all having distinct recording locations. For each recording location, 3–5 minute long audio recordings were captured. The original recordings were then split into segments with a length of 10 s. The total number of recordings were 4680, sampled at 44.1 kHz and were split into four folds (75/25 train/validation split). Given that our final goal was to train the network architectures on the DCASE 2017 development dataset and test it on the Extrasensory dataset (Vaizman et al. 2017), the MFCCs were extracted. The MFCCs were selected given that they are the most common features used in the fields of speech recognition and environmental sound recognition, and also since they are provided by the Extrasensory authors. The raw audio was not provided due to privacy issues. For the case of the MFCC feature extraction, the default sampling rate (44.1 kHz) of the DCASE dataset was used.

The number of MFCCs was 13 (including the $0th$ coefficient), the Fast Fourier Transform (FFT) window size was 2048, with a hop length of 1024 (50% overlap). This resulted in a 13×431 matrix per recording. Additionally, the mean and standard deviation of the MFCCs were calculated for each recording, resulting in a 13×1 vector. The main reason for selecting the aforementioned FFT window parameters, was that to match the same features extracted for the Extrasensory dataset. Therefore, a 1D CNN was used as the core architecture. The architecture was kept simple, based on the nature of the input data (mean and standard deviation of each MFCC). Three 1D convolutional layers were used, each followed by a max-pooling operation. The first layer had $f = 32$ filters, the second 48 and the third 120. The kernel size was set to $k = 2$. Each convolutional layer used the ReLU (Nair and Hinton 2010) activation function and the Adam optimizer was used, with an initial learning rate of 0.01. The network was set to train for 100 epochs and an early-stopping function was set that would stop the training if the validation loss was not improved after seven consecutive epochs. As with the IMU sensor, the weights of the CNN were frozen after the last max-pooling operation.

For these experiments two cases were considered. The first one has the aforementioned filter sizes and for the second case, we set the filter size to be 32 across all convolutional operations.

4.2 Step 2: Evaluation on real-world data

Evaluation in real-world conditions of the CNN feature extractors trained in the previous step was conducted using the Extrasensory dataset (Vaizman et al. 2017), including data from the smartphone inertial sensors (accelerometer and gyroscope) and audio recorded with the embedded microphone.

4.2.1 IMU case

Final classification on the Extrasensory was performed training an RF model taking auto-CNN features as input. At this stage, the optimization of hyperparameters of the RF model was performed. First, using a random search, with different hyperparameters (including number of estimators, max depth of each tree etc.).[2] Results were used to narrow the search, and to perform a grid search on a restricted number of combinations. The resulting model was used to simulate use of the pre-trained CNN feature extractor in a real-world scenario. This final validation (step 2 of the case study) was performed using the 5-fold validation provided with the Extrasensory (using 48 participants as test and 12 as validation).[3] The test on the Extrasensory data is much more challenging than the UCI-HAR. Given that the Extrasensory dataset was recorded in free-living conditions, the position of the smartphone is not constrained, which introduces further variability due to users having different habits, for instance between users keeping the smartphone in their trouser pocket or in their bag. The set of target activities is not the same. In this test, the aim was to use the CNN features to detect: lying, sitting, walking, running and cycling. A set of target activities that allows comparison with Vaizman et al. (2017) on the same dataset. The cycling class is typically more problematic, since it has often been reported as conflicting with the walking class (Incel et al. 2013). That is the case especially for users keeping the smartphone in their trouser pocket, where walking and cycling can end up generating similar patterns in the signal. The Extrasensory dataset is also representative of multiple devices, and mobile operating systems (Android and iOS). Finally, the Extrasensory dataset is highly imbalanced; a characteristic that is quite common in real-world datasets where balance between classes is not guaranteed, with respect to the case, for instance, of collecting data following a script. Class imbalance was addressed at the training stage by balancing the classes using undersampling, i.e., using random elimination of samples from the majority classes (lying and sitting). In the final evaluation, as in Vaizman et al. (2017) balanced accuracy, defined as macro-average recall (Pedregosa et al. 2011), was used to evaluate results, given that simple accuracy (ratio between correct and wrong classifications) can be biased in highly-imbalanced datasets. Evaluation on the Extrasensory was performed using the 5-fold partition provided with the dataset using 48 participants as training and

[2] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[3] Out of the 60 users of the extrasensory dataset, for 3 subjects no gyroscope data are available. The test was therefore limited to the remaining 57.

12 as test at each fold. Note, that in both cases (Extrasensory and UCI-HAR) the split into train test sets was performed using different subjects for training and testing, thus guaranteeing no overlap between the training and test set.

4.2.2 Audio case

Regarding the audio data, we focused on the classification performance of the 1D CNN network in an “unseen”, during training, dataset. The hosts of the Extrasensory dataset provided the 13 MFCC features extracted per frame. The dataset, however, contained recordings where the feature vector per recording was larger than the $(\approx 430) \times 13$, as described in the original paper (Vaizman 2017). Therefore, we worked with the 13 mean and 13 standard deviation MFCCs that were extracted per recording.

The target classes used for the experiment were related to the location of the user (indoor, outdoor and vehicle environment). As in the IMU case, it was assumed that the audio could be distorted based on the placement of the smartphone, e.g., in the pocket of the user. Specifically, regarding the indoor environments we selected the classes as seen in Table 1.

Since, our goal was to have the same types of classes with the pre-trained network on the DCASE 2017 development dataset, one could argue that the grouping of the Extrasensory classes is quite arbitrary. Nevertheless, since we do not have access to raw audio, classes such as ‘AT A PARTY’ could be either in an indoor or outdoor environment. Regarding the Extrasensory dataset, two users out of the 60 did not have audio data and were therefore eliminated from our experiments. For the evaluation, we used the provided 5-fold setup.

4.3 Environment

All experiments were conducted using Keras (Chollet et al. 2015), with TensorFlow (Abadi et al. 2015) as back-end. The RF model and evaluation metrics were implemented using sklearn (Pedregosa et al. 2011). The python source code of the project is available as git repository (Cruciani et al. 2019c).

5 Results

This section reports results obtained for the IMU and audio case. Results are reported for each case, firstly presenting the training of the CNN feature extractor (step 1 of our case study), and then presenting results of the tests performed on the Extrasensory dataset, evaluating the pre-trained CNN features in combination with an RF model, for the IMU case, and for the audio case without fine-tuning in realistic conditions. Results regarding the first step are complemented with some plots, visualizing class separation in the feature space, with varying configurations and hyperparameters. The visualization provides an additional insight that helped to analyse results as reported in the discussion Section.

5.1 IMU results

5.1.1 Step 1: IMU CNN feature extractor

As described in Sect. 4, the evaluation started testing an increasing number of layers, while maintaining a small number of filters, and small kernel size, refer to Fig. 5a. The figure shows the plots of the first three principal components obtained using PCA on the feature space defined by the features produced by the CNN. To improve readability of the figure, samples belonging to the lying class were excluded from the plot, since points belonging to that class were located far from all the other classes. Increasing the number layers n, the accuracy of models was observed to increase. The 3-CNN and the 4-CNN were the best performing models, with no significant difference between the two, despite the complexity added by the extra layer in the latter case. Consequently, the 3-CNN model was used to explore use of larger kernel sizes. Figure 5b shows the same plot maintaining the number of layers fixed and increasing the size of the kernel. The set of target activities in UCI-HAR consists of: lying, sitting, standing, walking, walking upstairs, walking downstairs (plus postural transitions). Two groups can be identified in this set: static activities (lying, sitting, standing) and dynamic activities. Larger kernel sizes of 24 and 32 (corresponding to approximately 0.5 s) were observed to improve discrimination between dynamic activities. This can be due to the periodic nature of the walking patterns, for which larger kernel sizes (able to capture at least the duration of a step) generate a more informative feature map, compared to smaller kernels.[4] Similarly, increasing the number of filters led to performance improvements up to $f = 24$. Further increase of the number of filters did not produce significant changes.

Figure 6 presents the confusion matrices obtained with the updated UCI-HAR dataset including transitions as NULL class. Table 2 presents the classification report obtained. These results were obtained using the 128×8 model, i.e., taking as input the three axes of accelerometer and gyroscope, and the 3D magnitude of the accelerometer and gyroscope signal.

[4] Note that the window size used for segmentation in the UCI-HAR dataset is of 128 samples, in order to capture at least a complete stride cycle of two steps (Anguita et al. 2013).

Table 1 Grouping of the extrasensory dataset classes

IN A MEETING, LOC main workplace,	INDOORS
SLEEPING, OR indoors, LOC home,	
IN CLASS, EATING, COOKING,	
LAB WORK, COMPUTER WORK,	
AT SCHOOL, SURFING THE INTERNET,	
WATCHING TV, DOING LAUNDRY,	
WASHING DISHES, CLEANING,	
FIX restaurant, AT A PARTY, ELEVATOR,	
TOILET	
OR outside, SHOPPING, LOC beach	OUTDOORS
BICYCLING, ON A BUS, IN A CAR,	VEHICLE
DRIVE-I AM THE PASSENGER,	
DRIVE-I AM THE DRIVER	

5.1.2 Step 2: IMU evaluation in realistic conditions

The CNN feature extractor trained in the previous step was then employed in the final step of the case study, evaluating its use in real-world settings on the Extrasensory dataset. Figure 7 presents the confusion matrix obtained using the RF model taking as input the auto CNN features. The set of target activities was set to lying, sitting, walking, running and cycling to allow comparison of results with Vaizman et al. (2017).

A simplified target activity set was also considered, combining the lying and sitting class into the *idle* state. The confusion matrix resulting from the 5-fold evaluation on the Extrasensory dataset is reported in Fig. 8.

Finally, Table 3 presents the classification report obtained for the 5-fold validation on the Extrasensory dataset.

5.2 Audio results

This section describes the results, using the 1D CNN architecture, that were obtained in the DCASE 2017 development dataset and the Extrasensory dataset.

5.2.1 Step 1: Training audio CNN feature extractor

Regarding the training of the CNN architecture, two case scenarios were examined and the precision, recall and F-score were calculated. The training and testing was performed on the DCASE 2017 development dataset, using the default 4-fold cross-validation setup.

We notice that while the difference in the macro-average of the metrics is very small (Tables 4 and 5), the PCA plot (Fig. 9) demonstrates that we can distinguish the outdoor classes with the vehicle and indoor. This means that when increasing the number of filters, the network can pay attention to more details in the signal, similarly to a

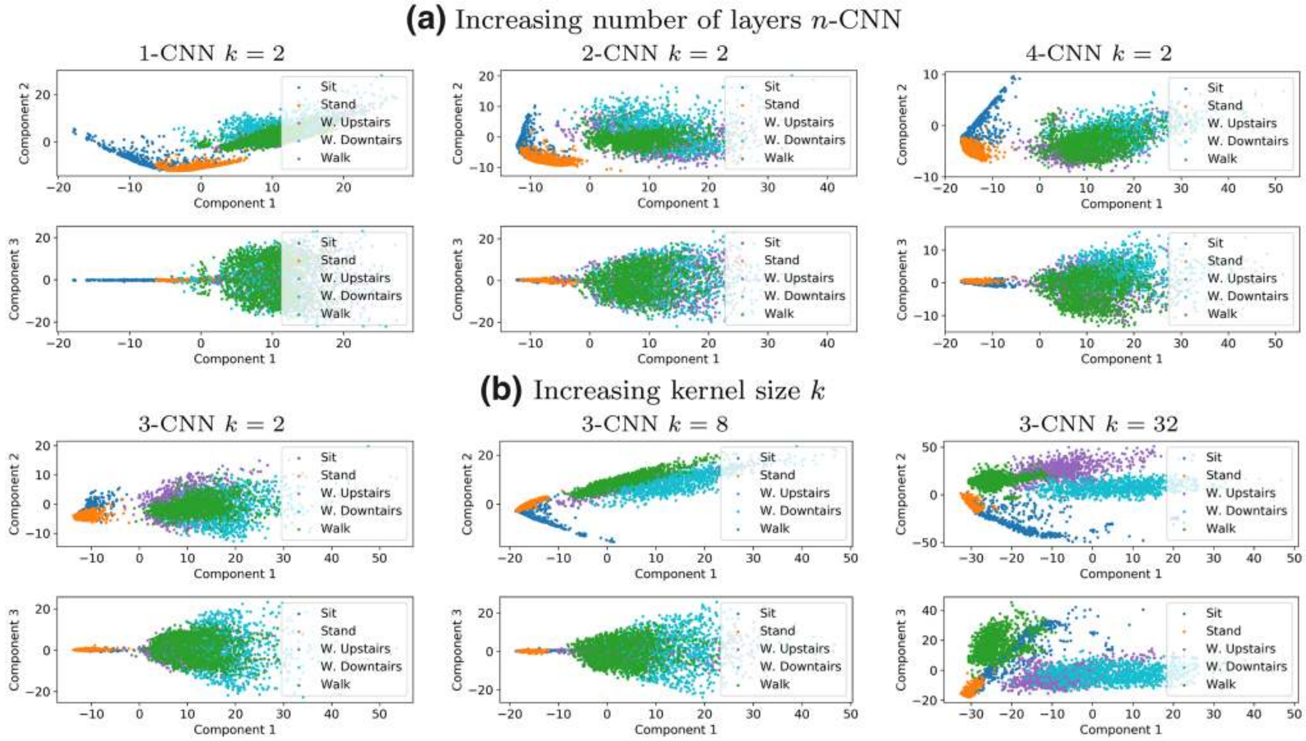


Fig. 5 PCA for the IMU sensor on feature space obtained by increasing the number of layers 1-CNN, 2-CNN, and 4-CNN (**a**); and **b** increasing the kernel size $k = 2$, $k = 8$ and $k = 32$

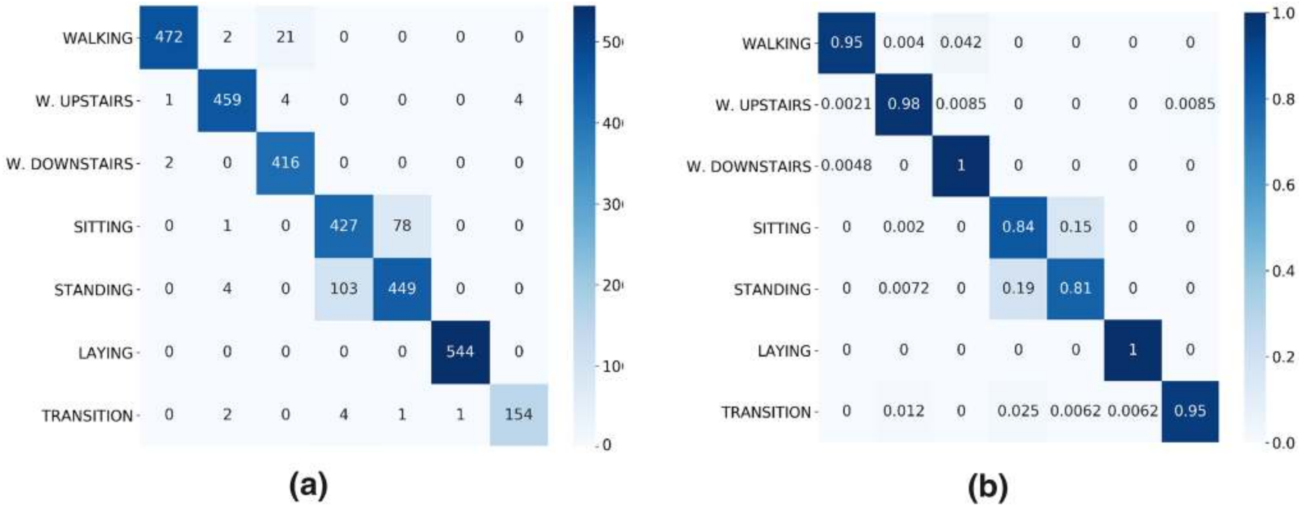


Fig. 6 Confusion matrix (**a**), and normalized confusion matrix (**b**) obtained with the final 3-CNN model with $k = 24$ and $f = 24$, for the set of target activities including transitions for the IMU case

Table 2 Precision, Recall and F-Score obtained with the 3-CNN ($k = 32$ and $f = 24$) and including transitions for the IMU case

Activity	Precision	Recall	F-Score
Walking	0.9979	0.9576	0.9773
W. Upstairs	0.9784	0.9701	0.9742
W. Downstairs	0.9265	0.9952	0.9596
Sitting	0.8867	0.7273	0.7991
Standing	0.7862	0.9191	0.8474
Lying	0.9963	1.0000	0.9982
Transition	0.9933	0.9198	0.9551
Average[a]	**0.9379**	**0.9198**	**0.9302**

The values in bold indicate the global average for all classes, to highlight the values as total results

[a] Macro average

computer vision problem, where the network learns finer details of an image.

5.2.2 Step 2: Audio evaluation in realistic conditions

The purpose of this experiment was to evaluate a pre-trained model on unseen data during training. The DCASE 2017 development dataset consists of 15 classes that were grouped in three main classes (indoor, outdoor and vehicle). On the other hand, the Extrasensory data is a much larger dataset, consisting of 52 classes. The dataset contains information not only about the activity of the user (e.g., walking), however, also about the location of the user (e.g., at a party). We noticed that there was not a one-to-one matching between the classes of the two datasets, despite the grouping into three classes. The results are summarized in Table 6. The indoor class had the largest precision and this is due to the fact that most of the classes were grouped as indoors, hence the imbalanced dataset.

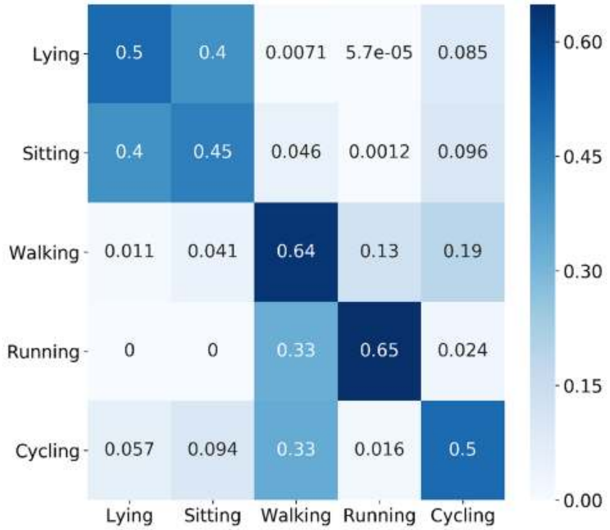


Fig. 7 Normalized confusion matrix of Fold 2 evaluating the RF model using the pre-trained CNN feature extractor for the IMU case

6 Discussion

The case study allowed the assessment of the effect of the main hyperparameters and CNN configurations on their feature learning abilities. The experiment provided a good overview of the main elements affecting feature learning abilities of a CNN for HAR. The results obtained in the first step of the case study highlighted how CNNs can perform at least as good as the best HCF, while providing a standardized

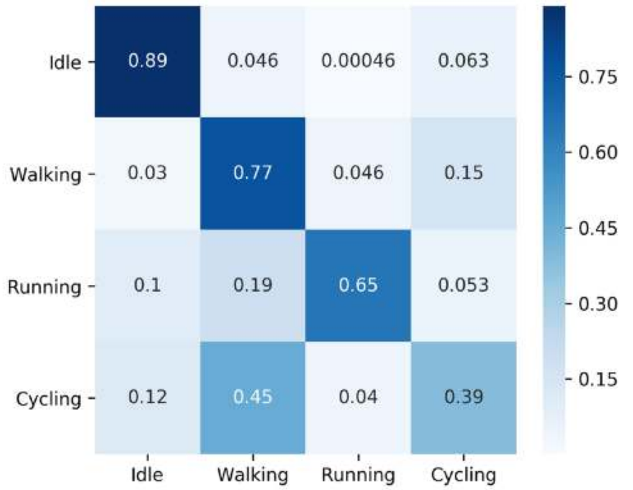


Fig. 8 5-Fold validation on extrasensory dataset using the Idle class for the IMU case

Table 3 Precision, Recall and F-Score obtained on the Extrasensory dataset

Activity	Precision	Recall	F-Score
Idle	0.9906	0.8907	0.9380
Walking	0.4330	0.7710	0.5546
Running	0.2328	0.6518	0.3430
Cycling	0.2135	0.3870	0.2752
Average[a]	**0.4675**	**0.6751**	**0.5277**

The values in bold indicate the global average for all classes, to highlight the values as total results

[a]Macro average

manner to accomplish the feature extraction step. On the other hand, the use of CNNs requires a training phase, making it subject to the cold-start problem. In this case study, CNN feature learning methods for audio and IMU cases were examined, and the use of a pre-trained CNN feature extractor was evaluated on a real-world dataset. The final evaluation on a real-world dataset allowed CNN automatic features to be tested under realistic circumstances. Overall, the test in realistic conditions highlighted the challenges of dealing with uncontrolled environments; for both the IMU and the audio case.

6.1 IMU

With respect to (Ronao and Cho 2016), where a similar analysis of CNN for HAR was examined, in this work, the more challenging case of a set of target activities including a NULL class was considered. While performing final training of the CNN feature extractor, it was noticed that SGD, although requiring a larger number of training epochs (compared to Adam), allowed training of the model with a higher accuracy on the test set, leading to 93.03% F-Score in the more challenging set of target activities including the transitions as NULL class. The analysis helped to identify a suitable CNN architecture providing required feature learning capabilities while aiming at keeping the complexity of the model under control. Building on results of the first step, a 3 layer CNN was identified and used as feature extractor in the second step. The architecture identified was tested on a large public available real world-dataset. Working with real-world datasets, as previously mentioned, introduces multiple variables that may affect accuracy performance. Such variables are usually under-represented in datasets collected in controlled environments. Despite the gap in accuracy measured in the first, and in the second step using real-world data, obtained results on the Extrasensory dataset using the pre-trained CNN were in line with (Vaizman et al. 2017); where HCFs were used on the same dataset, using the same input sensors (accelerometer and gyroscope), and with the same set of target activities. The measured balanced accuracy was 55.38% and 67.51% considering the idle (sitting/lying) class.

Table 4 Precision, Recall and F-Score obtained on the DCASE 2017 development dataset with the 3-CNN $k = 2$, and the same number of filters for the AUDIO case

Activity	Precision	Recall	F-Score
Indoor	0.9475	0.8811	0.9131
Outdoor	0.8808	0.9282	0.9039
Vehicle	0.8920	0.9263	0.9088
Average[a]	**0.9067**	**0.9119**	**0.9086**

The values in bold indicate the global average for all classes, to highlight the values as total results

[a]Macro average

Table 5 Precision, Recall and F-Score obtained on the DCASE 2017 development dataset with the 3-CNN $k = 2$, and increasing number of filters for the AUDIO case

Activity	Precision	Recall	F-Score
Indoor	0.9493	0.9151	0.9319
Outdoor	0.8897	0.9308	0.9098
Vehicle	0.9260	0.9231	0.9246
Average[a]	**0.9217**	**0.9230**	**0.9221**

The values in bold indicate the global average for all classes, to highlight the values as total results

[a]Macro average

6.2 Audio

Despite the reported recognition accuracy obtained for the audio case, we have shown that it is possible to perform a reasonable inference in an unseen environment, especially for the case of the indoor class. The performance of the network in the unseen training dataset, was affected by the

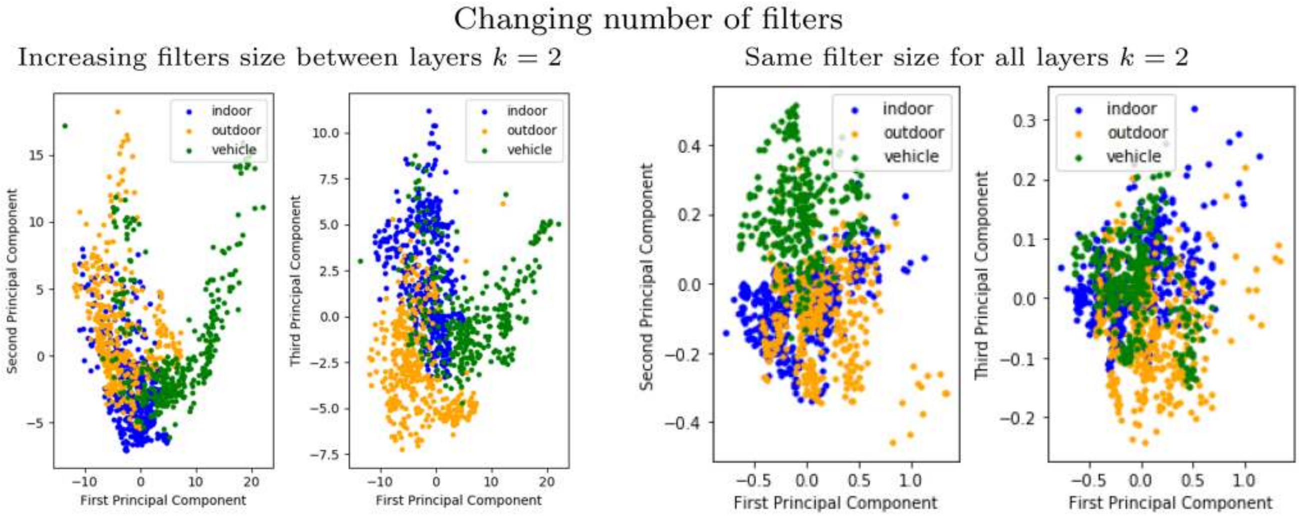


Fig. 9 PCA for the audio sensor on feature space obtained using the same (right plot) and different filter sizes between the layers (left plot)

Table 6 Precision, Recall and F-Score obtained on the Extrasensory dataset, for the 1D CNN trained on the DCASE 2017 development dataset for the AUDIO case

Activity	Precision	Recall	F-Score
Indoor	0.8928	0.2775	0.4235
Outdoor	0.0320	0.3840	0.0591
Vehicle	0.1004	0.3957	0.1602
Average[a]	**0.3417**	**0.3524**	**0.2142**

The values in bold indicate the global average for all classes, to highlight the values as total results

[a] Macro average

different sampling rate (44.1 kHz for the DCASE 2017 development dataset and 22 kHz for the Extrasensory dataset). The pre-trained network captured information that was present at higher frequencies, for instance at approximately 22 kHz, whereas the Extrasensory dataset could go up to 11 kHz (Nyquist theorem). Furthermore, the performance can be explained by the fact that the classes, where the DCASE dataset was grouped were not the same as in the Extrasensory dataset and the environments that were collected were much different, in terms of the acoustic conditions (reverberation, smartphone in pocket, smartphone's microphone quality). This justifies the problem that exists in the audio-based event recognition, where it is not possible to achieve a high recognition accuracy when testing in a new environment that contains different classes from the ones the algorithm was trained for. Therefore, there is a strong need for architectures that could be robust in terms of recognition accuracy, in an open set (data that have not been seen during the training).

7 Conclusion

We examined the use of CNN as feature learning method for HAR. Both IMU and audio-based HAR were considered. The experiments were conducted following a case study of two steps in which: (1) a CNN feature extractor is trained on a dataset collected in a controlled environment; subsequently, (2) the obtained pre-trained feature extractor is tested on a second real-world dataset evaluating its use as feature extractor in realistic condition. Results at all stages confirmed that CNNs can challenge the state-of-the-art HCF-based approaches, while providing a standardized and automated way to accomplish the feature learning step. At the same time, the use of a pre-trained CNN feature extractor can address the problem of the cold-start affecting CNN based approaches; although results obtained highlighted the multiple challenges of dealing with real-world data.

The goal of this work was primarily to provide an illustrative example of using a CNN pre-trained feature extractor, rather than providing a comprehensive analysis of all hyperparameters and configurations. Therefore, the optimization of models undertaken presents some limitations. For instance, while architectures with different numbers of convolutional layers and kernel sizes for the convolution were examined, only the ReLU activation function was used. Other activation functions may be considered in future studies. Similarly, the learning rate was kept to the Keras default value of 0.001.

Despite these limitations, the experiment provided a good overview of the use of CNN for HAR covering the effect of the main hyperparameters on discrimination of target activities on the feature space. As in more mature DL applications, such as computer vision, we can expect the use of pre-trained CNN models to become more common in the future. Pre-trained models can be used simply to initialize weights, or directly to extract features from raw data as in our case. Future work will include further investigation on CNN as feature learning method in order to develop reusable models for both IMU and audio-based HAR.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Abadi, M., Agarwal, A., et al.: TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/, software available from tensorflow.org (2015)

Abdel-Hamid, O., Ar, Mohamed, Jiang, H., Deng, L., Penn, G., Yu, D.: Convolutional neural networks for speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(10), 1533–1545 (2014)

Alsina-Pagès, R., Navarro, J., Alías, F., Hervás, M.: homesound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring. Sensors **17**(4), 854 (2017)

Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J.L.: A public domain dataset for human activity recognition using smartphones. In: 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN (2013)

Baldominos, A., Cervantes, A., Saez, Y., Isasi, P.: A comparison of machine learning and deep learning techniques for activity recognition using mobile devices. Sensors **19**(3), 521 (2019). https://doi.org/10.3390/s19030521

Bulling, A., Blanke, U., Schiele, B.: A tutorial on human activity recognition using body-worn inertial sensors. ACM Comput. Surv. (CSUR) **1**(June), 1–33 (2014)

Çakir, E., Virtanen, T.: End-to-end polyphonic sound event detection using convolutional recurrent neural networks with learned time-frequency representation input. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–7 (2018)

Chollet, F., et al.: Keras. (2015). https://keras.io

Cruciani, F., Sun, C., Zhang, S., Nugent, C., Li, C., Song, S., Cheng, C., Cleland, I., McCullagh, P.: A public domain dataset for human activity recognition in free-living. In: 2019 IEEE SmartWorld, 2nd SmarterAAL Workshop (2019a)

Cruciani, F., Vafeiadis, A., Nugent, C., Cleland, I., McCullagh, P., Votis, K., Giakoumis, D., Tzovaras, D., Chen, L., Hamzaoui, R.: Comparing CNN and human crafted features for human activity recognition. In: 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing (2019b)

Cruciani, F., Vafeiadis, A., et al.: Source code repository (2019c). https://github.com/fcruciani/cnn_rf_har

Eronen, A.J., Peltonen, V.T., Tuomi, J.T., Klapuri, A.P., Fagerlund, S., Sorsa, T., Lorho, G., Huopaniemi, J.: Audio-based context recognition. IEEE Trans Audio Speech Lang Process **14**(1), 321–329 (2006)

Espinilla, M., Medina, J., Salguero, A., Irvine, N., Donnelly, M., Cleland, I., Nugent, C.: Human Activity Recognition from the Acceleration Data of a Wearable Device. Which Features Are More Relevant by Activities? Proceedings vol. 2, no. 19, pp. 1242 (2018)

Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 776–780. IEEE (2017)

Grais, E.M., Wierstorf, H., Ward, D., Plumbley, M.D.: Multi-resolution fully convolutional neural networks for monaural audio source separation. In: International Conference on Latent Variable Analysis and Signal Separation, pp. 340–350. Springer (2018)

Huang, S.J., Gao, W., Zhou, Z.H.: Fast multi-instance multi-label learning. IEEE Trans Pattern Anal Mach Intell **41**(11), 2614–2627 (2019)

Incel, O.D., Kose, M., Ersoy, C.: A review and taxonomy of activity recognition on mobile phones. BioNanoScience **3**(2), 145–171 (2013)

Janidarmian, M., Fekr, A.R., Radecka, K., Zilic, Z.: A comprehensive analysis on wearable acceleration sensors in human activity recognition. Sensors **17**(3), 529 (2017)

Keskar, N.S., Socher, R.: Improving generalization performance by switching from adam to sgd. arXiv preprint arXiv:171207628 (2017)

Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference for Learning Representations (ICLR-15) (2015)

LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015). https://doi.org/10.1038/nature14539

Li, F., Shirahama, K., Nisar, M.A., Köping, L., Grzegorzek, M.: Comparison of feature learning methods for human activity recognition using wearable sensors. Sensors **18**(2), 1–22 (2018)

Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vincent, E., Raj, B., Virtanen, T.: Dcase 2017 challenge setup: Tasks, datasets and baseline system. In: DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events (2017)

Morales, J., Akopian, D.: Physical activity recognition by smartphones, a survey. Biocybern. Biomed. Eng. **37**(3), 388–400 (2017)

Morfi, V., Stowell, D.: Deep learning for audio event detection and tagging on low-resource datasets. Appl. Sci. **8**(8), 1397 (2018)

Moya Rueda, F., Grzeszick, R., Fink, G., Feldhorst, S., ten Hompel, M.: Convolutional neural networks for human activity recognition using body-worn sensors. Informatics **5**(2), 26 (2018). https://doi.org/10.3390/informatics5020026. http://www.mdpi.com/2227-9709/5/2/26

Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814 (2010)

Ordóñez, F.J., Roggen, D.: Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. Sensors **16**(1), 115 (2016)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)

Peltonen, V., Tuomi, J., Klapuri, A., Huopaniemi, J., Sorsa, T.: Computational auditory scene recognition. In: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 1941–1944 (2002)

Perttunen, M., Van Kleek, M., Lassila, O., Riekki, J.: Auditory context recognition using SVMs. In: Mobile Ubiquitous Computing, Systems, Services and Technologies, 2008. UBICOMM'08, IEEE, pp. 102–108 (2008)

Rajaraman, S., Antani, S.K., Poostchi, M., Silamut, K., Hossain, M.A., Maude, R.J., Jaeger, S., Thoma, G.R.: Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. PeerJ **6**, e4568 (2018)

Reyes-Ortiz, J.L., Oneto, L., Samà, A., Parra, X., Anguita, D.: Transition-aware human activity recognition using smartphones. Neurocomputing **171**, 754–767 (2016)

Ronao, C.A., Cho, S.B.: Human activity recognition with smartphone sensors using deep learning neural networks. Expert Syst. Appl. **59**, 235–244 (2016)

Saeed, A., Ozcelebi, T., Trajanovski, S., Lukkien, J.: Learning behavioral context recognition with multi-stream temporal convolutional networks. arXiv preprint arXiv:180808766 (2018)

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)

Vaizman, Y.: Context recognition in-the-wild: unified model for multimodal sensors and multi-label classification. PACM Interact. Mob. Wearable Ubiquitous Technol. **1**(1), 1–22 (2017). https://doi.org/10.1145/3161192

Vaizman, Y., Ellis, K., Lanckriet, G.: Recognizing detailed human context in the wild from smartphones and smartwatches. IEEE Pervasive Comput. **16**(4), 62–74 (2017). https://doi.org/10.1109/MPRV.2017.3971131. arXiv:1609.06354

Valero, X., Alias, F.: Gammatone cepstral coefficients: biologically inspired features for non-speech audio classification. IEEE Trans. Multimedia **14**(6), 1684–1689 (2012)

Xia, X., Togneri, R., Sohel, F., Huang, D.: Random forest classification based acoustic event detection utilizing contextual-information and bottleneck features. Pattern Recognit. **81**, 1–13 (2018)

Zhao, X., Wang, D.: Analyzing noise robustness of MFCC and GFCC features in speaker identification. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7204–7208. IEEE (2013)

Federico Cruciani received his BScEng and MScEng in Computer Science Engineering at University of Florence, Italy. From 2006 to 2016 he worked as software project manager for R&D at I+ and Orthokey focusing on Computer Assisted Surgery and eHealth, working on optical and inertial sensors applied to kinematic analysis and rehabilitation. In 2016, he joined Ulster University as Marie Curie fellow in Computer Science, starting his PhD investigation on Activity Recognition. He is currently working as a Research Fellow at Ulster University.

Anastasios Vafeiadis is currently a Computer Science PhD student at De Montfort University. He received his Master of Science in Electrical & Computer Engineering from Northeastern University, Boston in 2014 and the Bachelor of Science in Electrical & Computer Engineering from Worcester Polytechnic Institute, Worcester in 2012. He is also working as a research assistant at the Information Technologies Institute of CERTH. His research interests include audio signal processing, machine learning and acoustics.

Chris Nugent (M'96) received the B.Eng. degree in electronic systems and the D.Phil. degree in biomedical engineering from Ulster University, U.K., in 1995 and 1998, respectively. He is currently a Professor of biomedical engineering in the School of Computing, Ulster University. His research interests include data analytics for smart environments and the design and evaluation of pervasive and mobile solutions within the context of ambient-assisted living.

Ian Cleland received the B.Sc. degree in Biomedical Engineering and the PhD degree from Ulster University, U.K. He is currently a Lecturer in Data Analytics, within the School of Computing at Ulster University. His research focuses on the development and evaluation of novel healthcare technologies that incorporate concepts from pervasive computing, biomedical engineering and behavioural science.

Paul McCullagh received a BSc (1979) and a PhD (1983) in Electrical Engineering from Queen's University of Belfast. He is a Reader in Computing at Ulster University. His research interests include Biomedical Signal and Image Processing, Data Mining, Brain-computer interface, and Assisted Living applications.

Konstantinos Votis is a Senior Researcher (Researcher C') at the Information Technologies Institute of the Center for Research and Technology Hellas. He received the Diploma degree in Computer Engineering and Informatics and the PhD in Service Oriented Architecture and Semantic Interoperability between heterogeneous systems from the University of Patras, Computer Engineering and Informatics in 2002 and 2011 respectively. He also holds an MSc in Computer Science and Engineering (2004) and an MBA from the Department of Business Administration, University of Patras (2007). His research is related with data integration, SoA, Web accessibility, Semantic Web and ontologies, telemedicine, etc. He has authored numerous publications (more than 45) for international journals and international conferences, edited books and events. Since 2001, he has been involved in a large number of research and development projects funded by EC (e.g. COG, KWFGRID, ACCESSIBLE, AEGIS, VERITAS, APSIS4ALL, ATIS-4ALL, CLOUD4ALL, etc.) and the Greek secretariat of Research and Technology.

Dimitrios Giakoumis has been a post-doctoral research fellow at the Information Technologies Institute of the Center for Research and Technology Hellas since February 2007. He received his PhD in Electrical and Computer Engineering from the Aristotle University of Thessaloniki. His main research interests include affective computing, biosignals processing, emotion recognition and affective modeling, affect-related activity recognition, virtual & mixed reality, web services and pervasive computing.

Dimitrios Tzovaras is a Senior Researcher (Researcher A') and the Director of the Information Technologies Institute (ITI) at the Center for Research and Technology Hellas (CERTH). His main research interests include visual analytics, 3D object recognition, search and retrieval, behavioral biometrics, assistive technologies, information and knowledge management, multimodal interfaces, computer graphics and virtual reality. He has been working as a Researcher since September 1999 and he has been involved in more than 60 projects, funded by the EC and the Greek Ministry of Research and Technology. His involvement with those research areas has led to the co-authoring of over 80 articles in refereed journals and more than 150 papers in international conferences. He has served as a regular reviewer for a number of international journals and conferences.

Liming Chen is Professor of Data Analytics of the School of Computing at Ulster University, United Kingdom. He received his B.Eng and M.Eng from Beijing Institute of Technology (BIT), Beijing, China, and his Ph.D in Artificial Intelligence from De Montfort University, UK. His research interests include pervasive computing, data analytics, artificial intelligence, activity and behaviour analysis, user-centred intelligent systems and their application for digital health and Ambient Assisted Living (AAL).

Raouf Hamzaoui received the MSc degree in mathematics from the University of Montreal, Canada, in 1993, the Dr.rer.nat. degree from the University of Freiburg, Germany, in 1997 and the Habilitation degree in computer science from the University of Konstanz, Germany, in 2004. He was an Assistant Professor with the Department of Computer Science of the University of Leipzig, Germany and with the Department of Computer and Information Science of the University of Konstanz. In September 2006, he joined De Montfort University where he is a Professor in Media Technology. His research interests include image and video coding, multimedia communication systems, channel coding, pattern recognition and error control systems.