Signal, Image and Video Processing (2018) 12:1387–1394
https://doi.org/10.1007/s11760-018-1293-x

ORIGINAL PAPER

Novel approaches to human activity recognition based on accelerometer data

Artur Jordao[1] · Leonardo Antônio Borges Torres[2] · William Robson Schwartz[1]

Received: 21 December 2017 / Revised: 12 March 2018 / Accepted: 25 April 2018 / Published online: 9 May 2018

Abstract
An increasing number of works have investigated the use of convolutional neural network (ConvNets) approaches to perform human activity recognition (HAR) based on wearable sensor data. These approaches present state-of-the-art results in HAR, outperforming traditional approaches, such as handcrafted methods and 1D convolutions. Motivated by this, in this work we propose a set of methods to enhance ConvNets for HAR. First, we propose a data augmentation which enables the ConvNets to learn more adequately the patterns of the signal. Second, we exploit the attitude estimation of the accelerometer data to devise a set of novel feature descriptors which allow the ConvNets to better discriminate the activities. Finally, we propose a novel ConvNet architecture to explore the patterns among the accelerometer axes throughout the layers that compose the network. We demonstrate that this is a simpler way of improving the activity recognition instead of proposing more complex architectures, serving as direction to future works with the purpose of building ConvNets architectures. The experimental results show that our proposed methods achieve notable improvements and outperform existing state-of-the-art methods.

Keywords Human activity recognition · Accelerometer data · Attitude estimation features · Convolutional neural networks.

1 Introduction

In the past decade, human activity recognition (HAR) has been an active research topic, mostly because of its direct applications in person identification [10], health care [14], homeland security and smart environments [6]. For this purpose, sensor-based data have been widely explored due to their easy acquisition and fast processing in dedicated wearable sensors [1,16,21]. Recent technological advances have allowed the employment of smartphones and smartwatches to perform HAR, since these devices provide inertial sensors such as accelerometers, gyroscopes and barometers. Among these sensors, accelerometers have been used extensively since its signals have shown to represent well different categories of human activities [28,32].

The conventional paradigm to perform HAR describes the activity from inertial sensor data using handcrafted features and present these features to a classifier. However, recent works have presented better results employing convolutional neural network (ConvNets) as an alternative to handcrafted approaches [13,24,31]. The most recent studies that explore the idea of using ConvNets in HAR based on wearable sensors are the works of [11,12,15]. These works conduct an extensive analysis regarding the configurations of a ConvNet, aiming to construct an architecture able to achieve a high activity recognition rate. On the other hand, their proposed ConvNets architectures are not suitable, since they capture a small temporal pattern besides being sensitive to noise by data acquisition, mainly due to convolutional kernel design. Consequently, the recognition rate achieved by ConvNets might be compromised. Different from these works, instead of proposing complex architectures, we show that an adequate convolutional kernel can be a simpler way of improving the activity recognition results.

Another line of research to improve HAR is to use handcrafted features as input for the ConvNet. As shown in the work of Simonyan and Zisserman [29], handcrafted features can improve the learning of the ConvNets, by providing complementary clues able to better discriminate the categories of

✉ Artur Jordao
arturjordao@dcc.ufmg.br

[1] Smart Surveillance Interest Group, Computer Science Department, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[2] Electronics Engineering Department, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

activities. Similar ideas were employed by Jiang and Yin [15], where the result of a discrete Fourier transform, applied on the raw signal, is presented for the ConvNet.

Despite recent efforts to build suitable neural networks [5,11,12] and to design handcrafted features [3] (or perform combination of the two approaches [15]), to the best of our knowledge, no significant improvements have been achieved up to date. To this end, we propose a set of approaches to explore the essence of the accelerometer signal, providing a richer data representation which enhances the activity recognition. In summary, our work presents the following contributions:

1. A simple and accurate method to improve activity representation. The method consists in dividing each activity sample into blocks and then concatenating these blocks at the end of the signal before presenting it to the ConvNet.
2. A novel set of features descriptors, referred to as *attitude estimation features*, to describe the activities from accelerometer data. We show that these features enhance the data representation. Moreover, when combined with the raw accelerometer data, they provide a powerful clue to discriminate the activities.
3. A novel ConvNet architecture to explore the patterns among the accelerometer signal per axis throughout the layers that compose the network. Different from previous works that investigate many network parameters [5,11,12,15], we focus specifically on examining the convolutional kernel size. Therefore, our insights can help as direction to future works, with the purpose of building ConvNets architectures.

To validate the above claims, we use a recent human activity dataset based on sensorial data, the WHARF dataset [2]. On this dataset, our method achieves an activity recognition rate of 79.30%, outperforming existing state-of-the-art methods.

The remainder of this work is structured as follows. Section 2 reviews the state-of-the-art methods in HAR associated with inertial data. Next, Sect. 3 explains the proposed ConvNet architecture, partition technique and the attitude estimation features, respectively. Finally, Sects. 4 and 5 present our experimental results and concluding remarks, respectively.

2 Related works

The process of recognizing activities from wearable sensor data can be divided into two main classes, handcrafted and ConvNet-based methods. In this section, we present an overview regarding both classes of approaches.

To achieve a secure and reliable authentication system, Gafurov et al. [8] proposed an authentication system based on accelerometer data captured from a person's foot. In their study, the authentication is performed by detecting the best matching between the accelerometer signal cycles from training samples and the testing subject. Their results suggest that accelerometer data are able to provide an accurate user authentication mechanism. Additionally, the work of Charvatova et al. [4] reinforces evidence that wearable sensor data can be useful in different applications domains. Their work focuses on the classification of cycling activities, cycling up and down, from the global positioning system (GPS) and heart rate data. To this end, Charvatova et al. [4] evaluated different classifiers such as Bayesian, radial basis neural networks and k-nearest neighbors to recognize these activities. Different from [4,8], Hegde et al. [14] proposed to interpret the signal as an image and then compute Radon transform and Euclidean distance, on this image, to obtain a feature vector. Next, at the authentication stage, the correlation coefficient between the training and testing feature vectors is calculated to verify the authenticity of a person.

The aforementioned approaches employ the raw signal to feed the classifiers, but several works seek for accurate features to describe the raw signal before performing the classification. For this purpose, Kwapisz et al. [20] extracted average, standard deviation, time between peaks and binned distribution. Then, these features were presented to different classifiers to determine which activity they belong to. The authors evaluated various classifiers, such as decision tree (J48), logistic regression and multilayer perceptron, where the last one presented the best results. Catal et al. [3] proposed to combine the classifiers used by Kwapisz et al. [20] to enhance their results. To this end, Catal et al. [3] applied ensemble techniques on the same features suggested by [20], and consequently, they achieved a higher recognition rate.

Although handcrafted features have been widely employed to describe activities, an increasing number of works have employed deep learning (more specifically ConvNets) to learn features [5,11,15,31]. Deep learning is an end-to-end learning process that has achieved state-of-the-art results for many tasks, such as activity recognition [7,22], face recognition [23,27] and object detection [9,19]. In this process, features and classifiers are learned simultaneously and the method is able to learn the best features for the data and task at hand.

A successful deep learning approach is the work of Chen and Xue [5]. In their work, Chen and Xue [5] proposed a sophisticated ConvNet to classify the different categories of activities from accelerometer data. Their ConvNet consists of three convolutional layers followed by a 2×1 max-pooling layer. Despite presenting good results, their method is limited to walking-based activities, such as walking upstairs, step walking and walking quickly. In contrast, Jiang and Yin [15] developed a ConvNet architecture with two convolutional layers followed by average-pooling. Similar to our proposed work, the authors suggested to enhance the representation of the activities before presenting them to the network. To this end, Jiang and Yin [15] apply a discrete Fourier transform on the raw signal and present the transformed signal to the ConvNet.

Similar to [15], Ha et al. [12] proposed a multimodal ConvNet which consists of two convolutional layers with kernels of 3 × 3 and 5 × 5, respectively. In the first layer, the filters learn to separate each different modality (i.e., accelerometer and gyroscope). Then, inspired by [12], Ha and Choi [11] proposed to learn the filters of all the layers in the ConvNet separately for each heterogeneous modality (e.g., accelerometer and gyroscope). Therefore, they were able to improve by 2.19 percentage points (p.p.) the results achieved in [12]. By their results, it is possible to note that several efforts are performed to achieve small improvements.

An interesting aspect in the work of [11] is that the authors have shown an experiment where 2D convolutions (ConvNets) are more suitable than 1D convolutions in the context of HAR based on wearable sensor data, though the latter has to learn fewer parameters. In addition, they demonstrated that ConvNets provide better results than hidden Markov models and hidden conditional random fields, which are traditional approaches to model temporal information.

Different from the previous studies, we focus on providing a low-cost and better data representation, which aids the ConvNet to discriminate the categories of activities.

3 Methodology

In this section, we first present the proposed ConvNet architecture. Then, we introduce the proposed partition technique and the attitude estimation features.

3.1 ConvNet architecture

ConvNets were designed to, given an input matrix, automatically learn features that optimize a loss function [5,11,31]. To generate this input matrix from wearable data, we employ the same process as in [5,11,12,15], which consists of segmenting the raw signal using a temporal window of t seconds. For more details regarding this procedure, we recommend [5,12,15,30].

After generating the input matrix (one column per accelerometer axis, and one row for each time step, such that the signals are inserted column-wise in the input matrix), the next step is to build an adequate architecture to extract patterns from the accelerometer data. Different from previous works, which focus on proposing complex ConvNets architectures [11,12,15], we believe that setting an adequate kernel dimension is the most important parameter to achieve a high recognition accuracy. For this purpose, we analyze different dimensions to compose the convolutional kernels. Notice that the kernel height is responsible for extracting temporal patterns while the kernel width extracts the correlation between neighboring axes. By performing this analysis, we design a ConvNet composed of convolutional 12 × 2 kernels, where this value of width is responsible to capture the association between two neighboring pairs of accelerometer axes.

In our ConvNet, after each convolutional layer, we apply a max-pooling kernel of 2 × 1 to enable the network to be robust to spatial shift. We also employ the softmax regression, after the last network layer, to address the multi-class problem, a commonly used approach [5,11,12,15,31]. For a given testing sample, the softmax regression assigns a probability value for each class and assigns the activity label according to the highest probability.

Our complete ConvNet architecture consists of three layers with 24, 48 and 32 convolutional filters, respectively. Figure 1 illustrates the proposed ConvNet architecture. This configuration produces a ConvNet with 72,884 parameters to be estimated during training. An important aspect of our ConvNet is that it has the same convolutional kernel shape for all layers.

3.2 Partition technique

Previous works, in the context of handcrafted features, have shown that dividing the activities into small sub-activities can provide slightly better results [17,18]. Inspired by this, we propose to employ a partition technique to ConvNets. The intuition behind this method is to introduce diversity in the samples. In this way, we provide more capability for ConvNets to learn different signal patterns, which could be latent in the original samples, producing a more robust model.

The proposed partition technique works as follows. Initially, for a given input matrix, we divide it row-wise into P blocks of the same size. After that, we generate a set of indexes, S, starting from 1 to P. Then, we select randomly an element i, $i \in S$, and concatenate block i at the end of the input matrix (therefore, artificially augmenting the number of rows). We ensure that each index is selected only once; therefore, this process is performed P times. Algorithm 1 describes the aforementioned process.

According to our experiments, we verified that performing this random split provides better results than selecting the elements following a sorted order.

3.3 Features based on attitude estimation

To further improve the results in HAR, it is important to find appropriate features that provide better separation between classes in the sense of capturing the specific characteristics

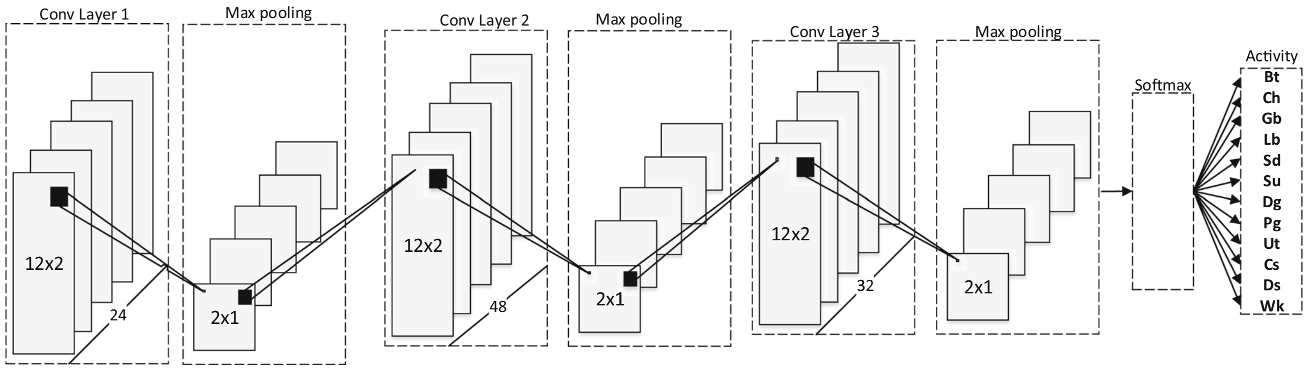

Fig. 1 Proposed convolutional neural network. All convolutional filters capture the temporal (by height) and the pattern among the accelerometer axes (by width)

```
Algorithm 1: Partition Technique.
Input  : X (input matrix with column-wise inserted signals)
Output : X concatenated with P shuffled temporal (row-wise) blocks
Segments = $X$ divided row-wise into $P$ blocks
$S = \{1, 2, 3, \ldots, |P|\}$
for $i = 1$ to $|P|$ do
    $i$ = random element of $S$
    $X = X + \text{Segments}_i$, (here $+$ denotes concatenation)
end
```

associated with each activity. In this work, from accelerometer data, we have estimated the wrist roll and pitch orientation angles, together with the three components of the wrist translational acceleration, which produces five time series used (i.e., five columns) in the input matrix to the ConvNet.

Notice that the measured acceleration can be represented as $\mathbf{a}_\mathrm{m} = \mathbf{a}_\mathrm{trans} - \mathbf{g} + {}^\circ$, where $\mathbf{a}_\mathrm{m} = [a_x\, a_y\, a_z]^\top$; $\mathbf{a}_\mathrm{trans}$ is the translational acceleration; $\mathbf{g}$ is the gravity acceleration vector, satisfying $\|g\| = g_0 = 9.806\,\mathrm{m/s}^2$; and ° represents additive zero mean measurement noise.

3.3.1 Pitch and roll

To estimate the wrist roll and pitch orientation angles, we have considered that their values change slowly in comparison with the accelerometer's sampling rate, for the considered activities. In addition, since the movement of a human arm is confined to a limited region in space, we have also assumed that high-frequency fluctuations in $\mathbf{a}_\mathrm{m}$ are mainly due to $\|{}^\circ\| \neq 0$ and therefore should be eliminated, since otherwise this would imply unrealistic arm movements. Based on these assumptions, we have performed the following two steps to estimate the wrist attitude angles: (i) low-pass filtering of the three raw accelerometer signals a_x, a_y and a_z to remove quantization errors and other additive high-frequency random noises, producing $\mathbf{a}^f_\mathrm{m} = [a^f_x\, a^f_y\, a^f_z]^\top$; and, (ii) by assuming at this point that the contribution of the wrist translational acceleration is much smaller than that of the local gravity, i.e., $\|\mathbf{a}_\mathrm{trans}\| \ll g_0$, and $\mathbf{a}^f_\mathrm{m} \approx -\mathbf{g}$; the wrist orientation was computed as

$$\phi = \operatorname{atan2}\left(-\, a^f_y \,,\, -\, a^f_z\right),$$
$$\theta = -\operatorname{asin}\left(-\, a^f_x \,,\, \sqrt{\left(a^f_x\right)^2 + \left(a^f_y\right)^2 + \left(a^f_z\right)^2}\right). \quad (1)$$

3.3.2 Translational acceleration

From the knowledge of θ and ϕ in (1), the gravity acceleration contribution $\mathbf{g}$ to the measured acceleration $\mathbf{a}_\mathrm{m}$ can be estimated as

$$\mathbf{g} \approx \mathbf{g}' = g_0\,[(-\ \sin\theta)\ (\cos\theta \sin\phi)\ (\cos\theta\cos\phi)]^\top,$$

and the estimated translational acceleration is given by

$$\mathbf{a}'_\mathrm{trans} = \mathbf{a}_\mathrm{m} + \mathbf{g}'. \quad (2)$$

By using $\mathbf{a}_\mathrm{m}$ instead of $\mathbf{a}^f_\mathrm{m}$, we guarantee that the whole spectral content of the original data is used. This is important because, if the cutoff frequency of the low-pass filter is not appropriately chosen and only the filtered data is used in the subsequent steps, valuable information might be lost, particularly for those activities associated with acceleration signals with high-frequency spectral content (e.g., *brushing teeth*).

3.3.3 Low-pass filtering

As we argued before, the high frequencies from accelerometer are generated by noise in the data acquisition; therefore, before computing the attitude estimation features, we need to eliminate them to avoid, for instance, unrealistic arm movements. For this purpose, we employed a fourth-order Butterworth low-pass filter with cutoff frequency ω_c =

$0.1\omega_s/2$, with ω_s the sampling frequency. For our data, in the majority of the cases, more than 90% of the energy of the accelerometer signals, computed after subtracting their time averages, was concentrated below ω_c. In addition, a zero-phase digital filtering [26] was implemented to guarantee the temporal synchronization between $\mathbf{a}_m(t)$ and $\mathbf{a}_m^f(t)$, and consequently between $\mathbf{a}_m(t)$ and $\mathbf{g}'(t)$, which is crucial to correctly compute $\mathbf{a}'_{trans}(t)$ from (2).

4 Experimental results

In this section, we first describe the dataset and the evaluation protocol used. Then, we present the experiments and results regarding our proposed ConvNet architecture. Afterwards, we demonstrate the gain provided by our attitude estimation features and the partition technique, respectively. Finally, we present experiments regarding the time issues, compare our method with state-of-the-art techniques and discuss some limitations of the attitude estimation features.

4.1 Dataset and evaluation protocol

To perform the activity recognition based on sensor data, we use the WHARF dataset [2]. This dataset consists of 12 activities, including brush own teeth (bt), comb own hair (ch), get up from the bed (gb), lie down on the bed (lb), sit down on a chair (sd), stand up from a chair (su), drink from a glass (dg), pour water into a glass (pg), use the telephone (ut), climb the stairs (cs), descend the stairs (ds) and walking (wk). The device, placed on the wrist of each voluntary, captures 3D accelerometer data using a sampling rate of 32 Hz.

We selected the WHARF dataset because it presents a large diversity in the activities and covers the most performed daily activities. Therefore, it is possible to examine the robustness of the methods regarding the high variance from the categories of activities to be classified. Additionally, the WHARF dataset provides samples acquired from low frequency, which makes the HAR a more challenging problem, as discussed in the study of Shoaib et al. [28].

To measure the accuracy achieved by the methods, we employed the widely used tenfold cross-validation protocol [3,11]. Finally, following [25,30], we segmented the raw signal using temporal windows of 5 s.

4.2 Convolutional neural network

In this experiment, we intend to show that our proposed ConvNet architecture is more appropriate than existing architectures, such as the ConvNet proposed by Chen and Xue [5], since it extracts the relation between the accelerometer axes (see Sect. 3.1) from the first to the last layer. To this end, we first evaluate the influence of the height and width dimensions which composes the convolutional kernel. As explained in Sect. 3.1, the kernel height is responsible for capturing the temporal relation while the width captures the association between pairs of accelerometer axes.

Table 1 Recognition accuracy (%) achieved by our proposed ConvNet using different kernel configurations

Kernel height	Kernel width	
	1	2
8	68.66	70.66
12	69.51	75.27
16	71.10	74.30

Table 2 Recognition accuracy (%) achieved by ConvNets using distinct features to generate the input matrix

Feature	Chen and Xue [5]	Proposed CNN
Raw signal (3)	72.29	75.27
Pitch and roll (2)	67.30	71.55
T. acceleration (3)	68.84	69.99
Concatenation (8)	74.88	76.24

The values within parentheses in the first column indicate the number of columns in the input matrix when using the respective feature. The last row represents the result employing the concatenation of the raw signal with our proposed features

As given in Table 1, our ConvNet achieves better results with the use of kernel width equals to 2 for any given height. This behavior takes place once the association learned by the convolutional filters (between the neighbors channels) generates a strong clue to discriminate the activities. Intuitively, increasing the kernel width would improve the results. However, the convolution and max-pooling processes decrease the input matrix size after each layer, preventing the use of larger kernels since the input matrix dimensions are small.

Based on these results, in the next experiments, we use the best configuration presented in Table 1 as architecture to compose our ConvNet, named *ProposedCNN*.

4.3 Attitude estimation features

In the experiment discussed above, we employed only the raw signal for generating the input matrix to the ConvNet. Now, we evaluate the impact of our proposed attitude estimation features on HAR. Table 2 presents the recognition accuracy achieved by ConvNet suggested by Chen and Xue [5] and our ConvNet, using different features to compose the input matrix. Notice that employing the raw signal, the ConvNets achieve slightly more accurate results regarding the use of our proposed features individually. However, when we combine all features (raw signal, pitch and roll, and translational acceleration), we achieve an improvement regarding the raw signal, for both our proposed ConvNet and the ConvNet proposed by Chen and Xue [5].

Table 3 Recognition accuracy (%) achieved by our proposed ConvNet using the partition technique

Feature	Proposed CNN	Proposed CNN^{+}	Improvement (p.p.)
Raw signal (3)	75.27	77.63	2.36
Pitch and roll (2)	71.55	73.54	1.99
T. acceleration (3)	69.99	75.93	5.94
Concatenation (8)	76.24	79.31	3.07

Proposed CNN^{+} denotes the proposed ConvNet when employing the partition technique to generate the novel input matrix. The values within parentheses in the first column indicate the number of columns in the input matrix when using the respective feature

Table 4 Time (in seconds) to compute the attitude estimation features and perform the forward step with each feature

Feature	Computation time	Prediction time
Raw signal	0	0.0098
Pitch and roll	0.0011	0.0105
T. acceleration	0.0012	0.0107
Concatenation	0.0022	0.0169

The second column shows the time to compute the respective feature descriptor. The third column denotes the time for our ConvNet to predict a testing sample using the respective feature

Table 5 Comparison with the state-of-the-art

Method	Recognition accuracy rate
Kwapisz et al. [20]	46.10
Catal et al. [3]	59.63
Jiang and Yin [15]	70.08
Chen and Xue [5]	72.29
ProposedCNN^{+} (Ours)	79.31

According to the results shown in Table 2, we conclude that the proposed features enable the ConvNets to learn more discriminative convolutional filters when coupled with the raw signal, providing a more accurate activity recognition.

4.4 Partition technique evaluation

The goal of this experiment is to show that the partition technique also provides improvements. Table 3 shows the recognition accuracy achieved by our ConvNet coupled with the proposed partition technique, on different features used to yield the input matrix. We refer to this technique as *Proposed CNN*$^{+}$. According to the results, the method is able to provide a significant improvement in the accuracy (see the last column in Table 3). However, in our experiments, we observed that the recognition accuracy saturates when we repeat the block i more than once (see Sect. 3.2). Therefore, we conclude that each block i should be inserted only once, yielding a new input matrix of height $2 \times |I|$, where $|I|$ is the original input matrix height.

4.5 Time issues

Since the proposed attitude estimation features demand an additional cost to be extracted, in this experiment, we aim at demonstrating that this computation cost has a slight influence on the activity recognition system. For this purpose, we measure the average time, by considering 30 executions, to extract the attitude estimation features. Moreover, with the addition of more descriptors, the prediction time (forward step) of the ConvNet increases. Therefore, we also measure the average time required to predict a testing sample.

Table 4 shows the average time (in seconds) to compute each attitude estimation features. Note that the raw signal is represented by zero value since there is no additional cost to estimated it. According to the results, it is possible to observe that the average time to estimate the pitch and roll, and translational acceleration features increases in 0.0011 and 0.0012 s, respectively. In addition, the prediction time using these features increases in 0.0007 and 0.0009 s, regarding the employment of raw signal. Therefore, the extra time added by our features is slightly superior to the raw signal, which enables its use in real systems.

4.6 Comparison with state-of-the-art techniques

This experiment compares our proposed method with published state-of-the-art methods [31]. For this purpose, we use our ProposedCNN^{+} (see Table 3, last row) since it employs all techniques proposed in this work.

Table 5 shows that our method outperforms current ConvNets dedicated to HAR based on wearable sensor data. Our ConvNet architecture achieves better results due to the convolutional kernel dimensions, which capture both temporal relation and association between pairs of accelerometer axes in all layers. Observe that our proposed ConvNet using any feature (see Table 2) outperforms, considerably, handcrafted features methods. In particular, according to the results in Table 5, we can conclude that ConvNet approaches can be more suitable than handcrafted features.

Table 6 Recognition accuracy obtained in each activity by our ConvNet using the raw signal (ProposedCNN) and the attitude estimation features (ProposedCNN$^+$)

	Activity bt	ch	gb	lb	sd	su	dg	pg	ut	cs	ds	wk
ProposedCNN	92.39	90.44	67.14	30.69	62.42	56.89	87.79	88.80	79.53	77.78	78.91	90.47
ProposedCNN+	92.96	91.63	70.69	40.59	71.34	65.27	90.33	89.87	81.87	82.91	79.69	94.54
Improvement (p.p.)	0.57	1.19	3.55	9.9	8.92	8.38	2.54	1.07	2.34	5.13	0.78	4.07

Brush own teeth (bt), comb own hair (ch), get up from the bed (gb), lie down on the bed (lb), sit down on a chair (sd), stand up from a chair (su), drink from a glass (dg), pour water into a glass (pg), use the telephone (ut), climb the stairs (cs), descend the stairs (ds) and walking (wk)

We have not compared our results with those reported in [11,12] since we have based our experiments in the WHARF dataset where multi-modalities are not considered.

4.7 Limitations of the attitude estimation features

Our last experiment focuses on the limitations of the attitude estimation features. To this end, we report the recognition rate obtained, for each activity, by our ConvNet using the raw signal (ProposedCNN) and our proposed methods (ProposeCNN$^+$).

As given in Table 6, even though we achieve notable improvements, our features do not increase the recognition rate for the activities *pour water into a glass, descend the stairs and brush own teeth*, abbreviated as pg, ds and bt, respectively. Intuitively, regarding the activity *descend the stairs*, this happens since it is performed with small arm movements; hence, our attitude estimation features are not able to produce discriminative features. On the other hand, to the activity climb the stairs, where a strong arm movement is required, our attitude estimation features are able to improve the recognition rate in 5.13 % points. Finally, the minor improvements regarding the activities pour water into a glass and brush own teeth are an effect of the temporal window, where few windows capture the arm movement.

We highlight that, except for these activities aforementioned, the proposed feature descriptors are able to achieve remarkable improvements.

5 Conclusions

This work presented a set of methods to improve the HAR from accelerometer data. First, we proposed a simple and efficient partition technique that improves the results in ConvNet-based approaches. Second, we developed a set of novel feature descriptors that exploit the attitude estimation of the accelerometer data. These features are able to improve the activity representation when combined with raw accelerometer data, allowing an accurate classification. Third, we proposed a novel ConvNet architecture and we demonstrated that determining a correct dimension to the convolutional kernels is a simple but effective way of achieving better activity recognition. Our methods are computationally efficient and are able to outperform existing state-of-the-art approaches, including ConvNets-based approaches and handcrafted features, on one of the most recent datasets to human activity recognition based on accelerometer data.

Acknowledgements The results presented in this paper were obtained through research on a project entitled “HAR-HEALTH: Reconhecimento de Atividades Humanas Associadas a Doenças Crônicas”, sponsored by Samsung Eletrônica da Amazônia Ltda. under the terms of Brazilian federal law No. 8.248/91.

References

1. Bayat, A., Pomplun, M., Tran, D.A.: A study on human activity recognition using accelerometer data from smartphones. In: MobiSPC’14, (2014)
2. Bruno, B., Mastrogiovanni, F., Sgorbissa, A.: Wearable inertial sensors: applications, challenges, and public test benches. IEEE Robot. Autom. Mag. **22**, 116–124 (2015)
3. Catal, C., Tufekci, S., Pirmit, E., Kocabag, G.: On the use of ensemble of classifiers for accelerometer-based activity recognition. Appl. Soft Comput. **37**, 1018–1022 (2015)
4. Charvátová, H., Procházka, A., Vaseghi, S., Vysata, O., Valis, M.: Gps-based analysis of physical activities using positioning and heart rate cycling data. SIViP **11**(2), 251–258 (2017)
5. Chen, Y., Xue, Y.: A deep learning approach to human activity recognition based on single accelerometer. In: IEEE SMC (2015)
6. Chua, J., Chang, Y.C., Lim, W.K.: A simple vision-based fall detection technique for indoor video surveillance. Signal, Image Video Process. **9**, 623–633 (2015)
7. Fernando, B., Anderson, P., Hutter, M., Gould, S.: Discriminative hierarchical rank pooling for activity recognition. In: CVPR (2016)
8. Gafurov, D., Bours, P., Snekkenes, E.: User authentication based on foot motion. Signal, Image Video Process. **5**(4), 457 (2011)
9. Girshick, R.: Fast R-CNN. In: International Conference on Computer Vision (ICCV) (2015)
10. Guven, G., Gürkan, H., Guz, U.: Biometric identification using fingertip electrocardiogram signals. Signal, Image Video Process. 1–8 (2018)
11. Ha, S., Choi, S.: Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In: IJCNN (2016)
12. Ha, S., Yun, J., Choi, S.: Multi-modal convolutional neural networks for activity recognition. In: IEEE SMC (2015)

13. Hammerla, N.Y., Halloran, S., Plötz, T.: Deep, convolutional, and recurrent models for human activity recognition using wearables. In: IJCAI (2016)
14. Hegde, C., Prabhu, H.R., Sagar, D.S., Shenoy, P.D., Venugopal, K.R., Patnaik, L.M.: Heartbeat biometrics for human authentication. Signal Image Video Process. **5**, 485 (2011)
15. Jiang, W., Yin, Z.: Human activity recognition using wearable sensors by deep convolutional neural networks. In: ACM Conference on Multimedia Conference (2015)
16. Karagiannaki, K., Panousopoulou, A., Tsakalides, P.: An online feature selection architecture for human activity recognition. In: IEEE ICASSP (2017)
17. Kim, H., Kim, M., Lee, S., Choi, Y.S.: An analysis of eating activities for automatic food type recognition. In: APSIPA (2012)
18. Kim, H.J., Choi, Y.S.: Eating activity recognition for health and wellness: a case study on asian eating style. In: IEEE (ICCE) (2013)
19. Kong, T., Yao, A., Chen, Y., Sun, F.: HyperNet: towards accurate region proposal generation and joint object detection. In: CVPR (2016)
20. Kwapisz, J.R., Weiss, G.M., Moore, S.: Activity recognition using cell phone accelerometers. SIGKDD Explor. **12**, 74–82 (2010)
21. Lee, J., Kim, J.: Energy-efficient real-time human activity recognition on smart mobile devices. Mob. Inf. Syst. **2016**, 12, Article ID 2316757 (2016)
22. Ma, M., Fan, H., Kitani, K.M.: Going deeper into first-person activity recognition. In: CVPR (2016)
23. Masi, I., Rawls, S., Medioni, G., Natarajan, P.: Pose-aware face recognition in the wild. In: CVPR (2016)
24. Morales, F.J.O., Roggen, D.: Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. Sensors **16**, 115 (2016)
25. Morris, D., Saponas, T.S., Guillory, A., Kelner, I.: Recofit: using a wearable sensor to find, recognize, and count repetitive exercises. In: CHI (2014)
26. Oppenheim, A., Schaffer, R.: Discrete-Time Signal Processing. Prentice-Hall, New Jersey (1989)
27. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: CVPR (2015)
28. Shoaib, M., Bosch, S., Incel, O., Scholten, H., Havinga, P.: A survey of online activity recognition using mobile phones. Sensors **15**, 2059–2085 (2015)
29. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS (2014)
30. Song, H., Thiagarajan, J.J., Sattigeri, P., Ramamurthy, K.N., Spanias, A.: A deep learning approach to multiple kernel fusion. In: ICASSP (2017)
31. Wang, J., Chen, Y., Hao, S., Peng, X., Hu, L.: Deep learning for sensor-based activity recognition: a survey. In: CoRR (2017)
32. Zeng, M., Nguyen, L.T., Yu, B., Mengshoel, O.J., Zhu, J., Wu, P., Zhang, J.: Convolutional neural networks for human activity recognition using mobile sensors. In: MobiCASE (2014)