Multiscale DCNN Ensemble Applied to Human Activity Recognition Based on Wearable Sensors

Jessica Sena, Jesimon Barreto Santos and William Robson Schwartz
Smart Surveillance Interest Group, Computer Science Department
Universidade Federal de Minas Gerais, Minas Gerais, Brazil

***Abstract*—Sensor-based Human Activity Recognition (HAR) provides valuable knowledge to many areas. Recently, wearable devices have gained space as a relevant source of data. However, there are two issues: large number of heterogeneous sensors available and the temporal nature of the sensor data. To handle those issues, we propose a multimodal approach that processes each sensor separately and, through an ensemble of Deep Convolution Neural Networks (DCNN), extracts information from multiple temporal scales of the sensor data. In this ensemble, we use a convolutional kernel with a different height for each DCNN. Considering that the number of rows in the sensor data reflects the data captured over time, each kernel height reflects a temporal scale from which we can extract patterns. Consequently, our approach is able to extract from simple movement patterns such as a wrist twist when picking up a spoon to complex movements such as the human gait. This multimodal and multi-temporal approach outperforms previous state-of-the-art works in seven important datasets using two different protocols. In addition, we demonstrate that the use of our proposed set of kernels improves sensor-based HAR in another multi-kernel approach, the widely employed inception network.**

***Index Terms*—Human Activity Recognition, Wearable sensors, Multimodal data, CNN Ensemble, Multiscale Temporal Data**

I. Introduction

The use of sensors from wearable devices to recognize human activities has grown every year. As discussed by Lara et al. [1], there are many reasons for this growth: the increasing interest of several areas, such as, medical, military, and security applications; the convenience and comfort of using such devices (it does not change or hinders the action due to their use); the feeling of privacy (as opposed to monitoring with cameras where depending on the activity performed or the location, the user feels uncomfortable); and it is already naturally inserted into people's lives, facilitating the data capture. The number of sensors in such devices is increasing and the large range of sensors provide rich and complementary information regarding the activities performed by users. Therefore, an important line of research that has gained attention focuses on the investigation to combine (i.e., fuse) these multiple sensors to improve human activity recognition.

Some works perform fusion in the raw data (i.e., early fusion), concatenating the sensors into a common matrix used as input for machine learning methods. For instance, Chen and Xue [2] employed a Deep Convolutional Neural Network (DCNN) with three convolutional layers and used the size of the kernel to extract the relation between the axes and temporal information. Motivated by the architecture proposed in [2], Jordao et al. [3] suggested a DCNN able to explore the patterns among the signal axes in all the layers that compose the network. As a consequence, their proposed DCNN achieved better results than [2]. Different from [2], [3], Jordao et al. [4] employed a DCNN and use partial least squares analysis to reduce the dimensionality of each max-pooling layer and consider the concatenation of the dimension reduction as a feature to feed a softmax classifier. To improve the data representation, Jiang and Yin [5] applied a discrete Fourier Transform to preprocess the input matrix and use a DCNN composed by a stack of two convolutional layers, a fully connected and a softmax layer to recognize the activities. However, due to the multimodal nature of each sensor, merging the sensors in the raw data may not be appropriate since sensors have several dissimilarities between them, such as a different number of axes, scales, meanings, or data nature (e.g., angle, value, degree, frequency).

To address the multimodality problem, some authors proposed to insert a padding between the sensors to separate the data and to be able to extract features from the sensors separately. For instance, Ha et. al. [6] preprocessed the matrix of sensors adding a zero-padding between each sensor and use a DCNN with the same layer structure as in [5]. However, this division is only effective at the first layer since, from the second layer onwards, the data from different sensors are convoluted together. In fact, in another work, Ha and Choi [7] proposed to insert zero-padding before each convolutional layer to avoid interference between sensors when 2D convolutional kernel is applied. While this approach separates in some way the data before performing fusion, it uses the same DCNN to learn features from all sensors simultaneously, which might overcharge the model since the kernel have to learn patterns from different data nature.

In a recent work, Yao et al. [8] brought a new perspective on merging multimodal data to perform sensor-based HAR. They build an architecture with three different sequential blocks: an individual deep convolutional subnet for each input sensor to learn local patterns, a common deep convolutional subnet that concatenates all sensors and learns the high-level relationship among them and, at the end of the architecture, a stacked Gated Recurrent Unit [9] structure to learn meaningful temporal features. Since the use of convolutional and recurrent networks are already well established in the sensor literature, the main advance of [8] is to go beyond just placing a boundary between the sensors in the input matrix. Instead, they separate the

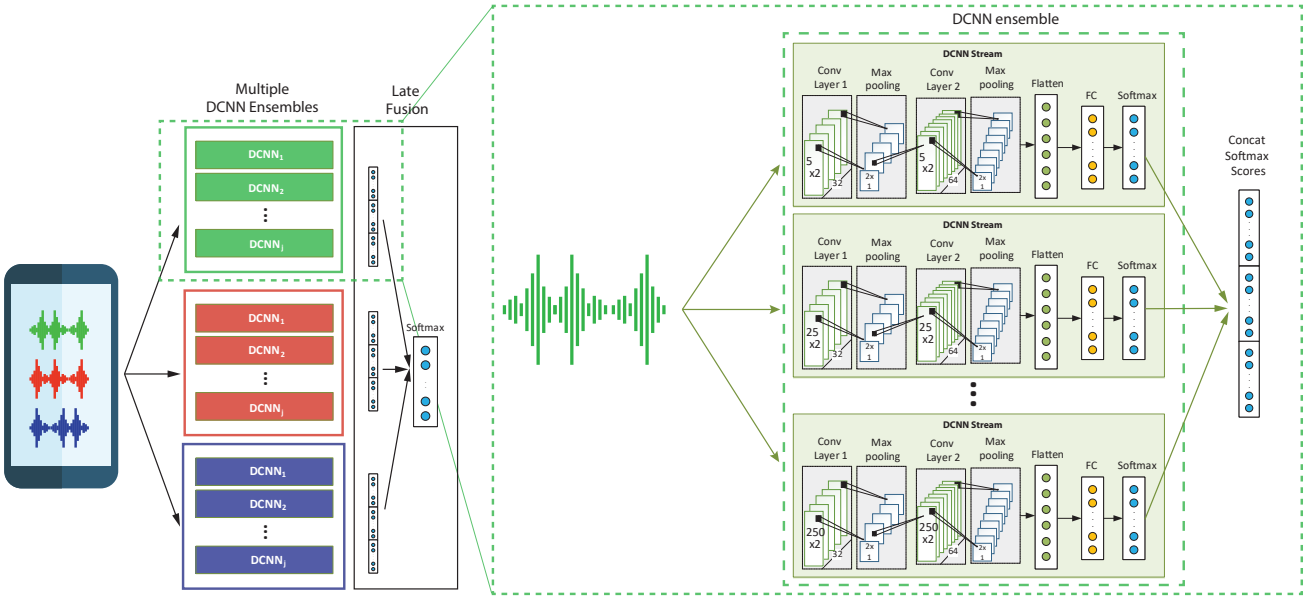

Fig. 1: Our Multimodal DCNN Ensemble (MDE) relies on two premises. The first is separately processing each sensor and the second is to extract patterns from multiple temporal scales. Thus, for each sensor, we create a DCNN ensemble that extracts multi-temporal information. This ensemble is composed of streams so that each one extracts patterns on a specific temporal scale and classifies the sample. We merge all scores into a late fusion approach which allows us to take advantage of the complementarity between both sensors and temporal scales.

sensors from the beginning to extract features individually and learn which patterns separate human activities for each sensor before merging and benefiting from their complementarity.

Besides the sensor data heterogeneity, another issue that must be considered is the temporal nature of the data. Due to the CNN input format for sensors (where columns refer to the sensor axes and rows to data-capture over time), the height of the convolutional kernel represents the size of the temporal window used to learn patterns. Since there are several possibilities to set the kernel height, we can see each size as a temporal scale to extract potential patterns.

In traditional deep convolutional network methods [2], [3], [5], a single kernel is set for each layer, which discards all other possible temporal scales for that particular layer. In these networks, each stacked convolutional layer learns features at a larger semantic level than the previous one and, in the sensor context, a deeper CNN network would learn features in multiple temporal scales due to its depth (each layer learns a higher temporal scale than the previous one). However, the convolutional maps that go to the next layer are the activations for the kernel in the previous layer. In this way, when one chooses a single kernel size for a specific layer, it might discard important information in this layer which would only be selected by another kernel size. Therefore, to avoid this problem, we propose the use of an ensemble of multiple kernels which is able to learn several temporal scales simultaneously. This follows the intuition that human activities are composed by different durations, i.e., while some activities can only be distinguished by small and fast movements, others need to be analyzed for longer periods of time to be classified.

Therefore, we propose an approach based on multiple streams to individually process the sensor data. Although, the core of this approach is a novel way to extract temporal data by employing an ensemble of temporal scales implemented with multiple DCNNs. As each DCNN has a kernel size which reflects one scale of a pre-defined temporal scale range, we can extract patterns of multiple sizes, ranging from short movements, such as a gentle twist of the wrist, to large and complex motions, such as the human gait. According to experimental results, our approach outperforms previous state-of-the-art results in seven datasets using two different evaluation protocols. In addition, we adapt the Inception module [10] to compare to our DCNN Ensemble approach (without multimodal premise) and we demonstrate that our method is better than the Inception. In addition, we show that using our kernel set is more suitable for the sensor-based HAR than the kernels originally proposed in the Inception module.

II. Proposed Approach

As illustrated in Figure 1, in our Multimodal DCNN Ensemble (MDE) approach, we first separate the sensors into different inputs to process each one individually. Then, for each sensor, we construct an ensemble of temporal scales extracted through DCNN streams that are subnets within our network. Finally, we use an approach based on late fusion to merge the multi-modal and multi-temporal information. This process is detailed in the following paragraphs.

DCNNs Ensemble. The sensor data is commonly stored in a matrix of size $t \times a$, where a is the number of axes of the sensor (for instance, 3 axes (x, y, z) on motion sensors) and t is the temporal axis, where each row is a sensor capture over time. Therefore, given a 2D kernel (h, w), our premise is that the

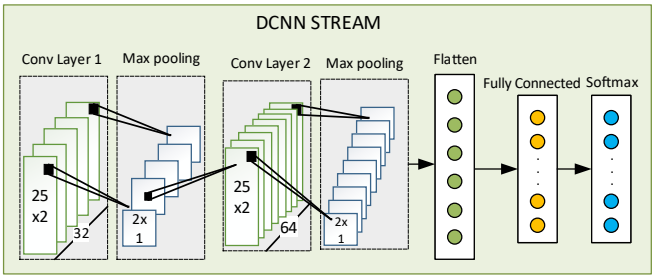

Fig. 2: The deep convolution neural network stream.

height of the kernel (h) is responsible for determining in which temporal scale we are learning the patterns. For instance, a h equal to 25 in a sample captured at a frequency of 100Hz learns patterns of 0.25 seconds while a h equal to 250 learns patterns of 2.5 seconds. Thus, the larger the kernel height, the larger the temporal pattern we capture.

To extract information from multiple temporal scales, we employ an ensemble of DCNNs with different kernel sizes each. As shown in Figure 1, an ensemble is built for each sensor, so we have several ensembles in our network, according to the number of sensors processed (i.e., in Figure 1, we have three sensors and consequently, three ensembles of DCNNs). The number of DCNNs in each ensemble is pre-defined as a parameter of our architecture called *pool*. The pool is a set of kernels $K = \{K_1, K_2, ..., K_j\}$ which contains a variety of kernel sizes ranging from a small kernel up to a large one. For each kernel in our pool, we add a DCNN stream in the ensemble and set its two convolutional layers with the specific kernel. For instance, in Figure 1, we have a pool of j kernels where three of them have their streams explicitly drawn in the figure composing a kernel pool $K = \{5\times2, 25\times2, ..., 250\times2\}$.

DCNN Stream. Each DCNN in the ensemble, for convenience, let us call it a *stream*, is a network composed of two parts, as shown in Figure 2. First, there is a convolutional block with two convolutional layers, intercalated by two max-pooling layers. The use of convolutional layers allows us to learn temporal patterns in the scale we define for each stream and the application of max-pooling controls overfitting, reduces the number of parameters and the computation cost. Second, at the end of the subnet, we have a fully connected block consisting of a fully connected and a softmax layer. We use scaled exponential linear units [11] as the activation function of the fully connected block. While the convolutional block provides a meaningful, low-dimensional, and somewhat invariant feature space, the fully-connected block is learning a non-linear function in that space, which translates the features extracted by the convolutional block to the softmax scores.

Late Fusion. After the previous stage, we have an ensemble for each sensor, and each ensemble outputs j probability vectors. It is necessary to merge this information to take advantage of the complementarity provided by both the multiple sensors and the multiple temporal scales. We empirically found that the best way to merge these streams is by using meta-learning of the scores. Thus, we concatenate all the score vectors of the streams ($j \times$ number of sensors) in a single feature vector and pass it to the classification layer (softmax). The training of our network is done in an end-to-end way, which optimizes the weights of the entire network since it maps the input of all the modalities to a single output. Consequently, the network dynamically learns which scales and sensors are most appropriate for each activity.

III. Experimental Results

One of the most latent problems in wearable sensor-based human activity recognition is the lack of standardization of metrics, evaluation protocols, and datasets, which makes it difficult a comparison among methods. While some works record their own datasets to perform experiments, others use datasets from the literature but do not clarify how to reproduce the experiments. Recently, a work has endeavored to solve this issue by bringing the first standardization to the domain. Jordão et al. [12] performed a thorough study and standardized seven datasets of the wearable sensor literature in different protocols. In this section, we quickly describe the experimental setup employed in this work using the framework proposed by [12] and then, we discuss the results achieved by our proposed approach and two simplifications of the approach.

A. Experimental Setup

Jordão et al. [12] conducted a survey in the literature and gathered seven important datasets: WHARF [13], USCHAD [14], UTD-MHAD (set 1 and 2) [15], WISDM [16], PAMAP2P [17] and MHEALTH [18]. This set of datasets composes an interesting diversity of number of samples, types of activities performed and number of available sensors, making it possible to evaluate the robustness of the methods in different scenarios. The datasets were processed and standardized with a sampling rate of 5 seconds, except for the UTD-MHAD dataset that had to be sampled at 1-second rate. We evaluate our approach in these seven datasets following strictly the procedure defined by Jordão et al. [12][1].

Regarding protocols, according to [12], Leave-One-Subject-Out (LOSO) and Leave-One-Trial-Out (LOTO) are the most appropriate for reporting results in sensor-based HAR. In the LOSO protocol, the data are separated in training and test so that the test has only one subject at a time and the training has the other subjects. In the LOTO, the trial consists of a transition from one activity to another, so the data is separated into trials where each trial contains only a continuous capture of an activity. Therefore, the training is performed with all the trials except one that is put to test. LOSO represents the real scenario of applications for wearables devices, where a method is trained in known subjects and applied to new subjects later. This protocol also analyzes the generalization quality of the method since the training and test data do not have the same distribution. On the other hand, LOTO protocol has the benefits of generating a large number of samples and certifying that the contents of a trial do not appear in training

[1]Refer to [12] for more details regarding the evaluation procedure.

and testing at the same time, different from the cross-validation protocols inappropriately used in the literature, which ensures a correct evaluation of the performance.

B. Results and Discussion

To evaluate separately the contribution of the DCNN ensemble and the multimodal hypothesis (processing each sensor separately), we implemented two simplified versions of our MDE method. In the first, called *DCNN Ensemble* we do not separate the sensors, instead, we concatenate all sensors into a single array (in the same way of the majority of works) that feeds one ensemble of kernels. In the second, called *Multimodal Stream*, we use only the multimodal hypothesis, using just one DCNN stream (see Figure 2) for each sensor instead of an ensemble. In this DCNN stream, we set a kernel of size 25×2 that empirically showed a good performance. Due to the multimodal architecture, we evaluate the MDE and Multimodal Stream only on datasets that contain more than one sensor.

We compare our approach with all methods evaluated by Jordão et al. [12]. Thereby, in addition to the methods mentioned in Section I, we also show results from three other handcrafted methods [19]–[21] surveyed by [12]. Usually, this family of methods extracts statistical features and applies a classifier to recognize activities. We include them in our evaluation mainly because they present better results in some datasets than the proposed approaches based on deep learning. Furthermore, we discuss more deeply the results of Yao et al. [8] in contrast to ours, since to the best of our knowledge, that is the only multimodal method using multiple streams that have been proposed so far in the context of wearables sensors. To analyze the contribution of the pool of kernels and to evaluate our DCNN ensemble, we use the Inception network module [10] as a baseline. Although the Inception was originally designed for object detection in images, it is analogous to our approach since it also applies multiple kernels to the same input to extract different pattern sizes.

Comparison with Kernel Ensemble Baseline. We could not compare our DCNN Ensemble with Inception's full architecture [10] because the available datasets do not have enough data to train a network of the size of Inception (in the object detection domain the Inception was trained using 1.2 million of images provided by ImageNet dataset [22], in our context, the dataset with the largest number of samples used in our evaluation has 20k samples). One option would be to use the pre-trained network on the ImageNet, by performing a transfer learning, but the pre-trained model is restricted to the use of three channels and to have a minimum array of 139×139 pixels. Besides to the sensory data being one channel, our largest dataset has a matrix of 500×10, so it is not possible to use the pre-trained Inception network. Therefore, we did a study of the appropriate number of Inception modules that should be used for the context of wearable sensors. The experiments showed that the addition of more than one module deteriorated the results, thus, all Inception-based experiments in this work were done by using only one Inception module.

Another important point is that we add to the Inception module the fully connected block used in our DCNN stream. This considerably increased the Inception performance, since the fully connected block is capable of fusing the different patterns extracted by the different kernels sizes and also regularize the network since we use SELU activation function. We employed as baselines the two modules proposed by Szegedy et al. [10]: the naïve and the dimensionality reduction module. In addition, to evaluate our kernel pool, we adapt each type of Inception module to the wearable sensors domain by using the same pool of kernels used by our approach instead of the kernels proposed in [10]. Table I shows the results obtained from these four approaches. It is possible to note that using the kernels pool improves the result of the Inception original modules for all datasets. This support our hypothesis that extracting multiple temporal scales is appropriate for the sensor domain. Besides, our DCNN Ensemble approach outperforms all four Inception-based methods using LOSO and LOTO on the seven datasets, which points out that our ensemble is more suitable to employ the use of multiple kernels to extract temporal information in the context of wearables sensors.

Comparison with Multimodal Baseline. Yao et al. [8] brought advances to sensor fusion with an approach based on multiple streams to processes each sensor separately. Our Multimodal Stream and MDE approaches follow this intuition. Table I shows that the results achieved by [8] are modest and in some cases smaller than very simple approaches like handcrafted methods. We believe this is because the network proposed by [8] has a very complex network which can cause overfitting since the datasets do not have a large number of samples. In addition, in the datasets of the UTD-MHAD family, the sample size does not allow it to be divided into time-steps to fed the network proposed in [8]. Thus, the approach performs poorly in the two UTD-MHAD datasets. Our work, on the other hand, showed superior results even using only the multimodal hypothesis through our Multimodal Stream approach (without DCNN Ensemble). Furthermore, using the MDE, we solve the temporality issue in an apparently more efficient way since it does not use recurrent networks and still surpasses more sophisticated approaches such as [8].

Comparison with the State-of-the-art. In the Table I is showed the results of our main approach, MDE, as well as two simplifications of it, the DCNN Ensemble and the Multimodal Stream (both explained at the beginning of this section). Our approaches overcome the results of our two baselines (Inception module [10] and Yao et al. [8]) and all methods of the literature surveyed by Jordão et al. [12] achieving, to the best of our knowledge, the state-of-the-art in the seven datasets evaluated. Particularly, in the datasets MHEALTH and PAMAP2P, the DCNN Ensemble approach showed superior results to the MDE approach in both protocols tested. We believe this is occurring because we had to reduce the number of parameters in MDE network for these two datasets due to the limited computational resources available to run our

	WHARF	UTD-1	UTD-2	WISDM	USCHAD	MHEALTH	PAMA	WHARF	UTD-1	UTD-2	WISDM	USCHAD	MHEALTH	PAMA
METHODS	LOTO (ACCURACY (%))							LOSO (ACCURACY (%))						
Kwapisz et al. [19]	44.51	15.99	69.61	79.08	76.52	89.75	70.58	42.19	13.04	66.67	75.31	70.15	90.41	71.27
Catal et al. [20]	64.84	47.80	81.37	80.52	87.77	91.84	81.03	46.84	32.45	74.67	74.96	75.89	94.66	85.25
Kim et al. [21]	61.12	50.98	75.27	56.26	85.70	91.51	78.08	51.48	38.05	64.60	50.22	64.20	93.90	78.08
Chen and Xue [2]	72.55	-	-	86.55	84.66	89.95	82.32	61.94	-	-	83.89	75.58	88.67	83.06
Jiang and Yin [5]	70.79	-	-	83.82	80.73	52.78	-	65.35	-	-	79.97	74.88	51.46	-
Ha et al. [6]	-	-	-	-	-	85.31	80.13	-	-	-	-	-	88.34	73.79
Ha and Choi [7]	-	-	-	-	-	82.75	71.19	-	-	-	-	-	84.23	74.21
Yao et al. [8]	×	12.70	22.41	×	81.34	31.35	70.59	×	11.45	22.40	×	71.52	31.88	72.61
Inception naïve mod [10]	43.98	50.87	76.27	83.02	-	-	-	36.64	40.71	72.55	78.64	-	-	-
Inception naïve + pool	49.86	53.06	76.71	84.89	-	-	-	41.14	41.44	72.46	81.99	-	-	-
Inception mod [10]	51.76	52.36	74.62	79.18	-	-	-	42.07	39.62	68.34	73.86	-	-	-
Inception + pool	60.74	56.66	78.62	86.83	-	-	-	49.97	42.23	72.96	80.99	-	-	-
DCNN Ensemble	**75.50**	62.03	81.63	**89.01**	88.49	**93.09**	**83.99**	**69.79**	46.75	79.38	**86.22**	82.66	**96.27**	**87.59**
Multimodal Stream	×	48.90	79.82	×	85.95	83.17	79.62	×	36.99	74.59	×	79.68	90.20	80.58
MDE	×	**69.61**	**83.78**	×	**90.08**	84.61	76.35	×	**57.13**	**81.99**	×	**83.40**	88.97	77.70