2017 IEEE International Conference on Data Mining Workshops

Towards automatic feature extraction for activity recognition from wearable sensors: a deep learning approach

Belkacem Chikhaoui[1,2]
[1] *Department of Science and Technology*
TELUQ University
Canada
belkacem.chikhaoui@teluq.ca

Frank Gouineau[2]
[2] *Computer Research Institute of Montreal*
Canada
Frank.Gouineau@crim.ca

***Abstract*—This paper presents a novel approach for activity recognition from accelerometer data. Existing approaches usually extract hand-crafted features that are used as input for classifiers. However, hand-crafted features are data dependent and could not be generalized for different application domains. To overcome these limitations, our approach relies on matrix factorization for dimensionality reduction and deep learning algorithm such as a stacked auto-encoder to automatically learn suitable features, which will be then fed into a softmax classifier for classification. Our approach has potential advantages over existing approaches in terms of automatic feature extraction and generalization across different application domains. The proposed approach is validated using extensive experiments on various publicly available datasets. We empirically demonstrate that our proposed approach accurately discriminates between human activities and performs better than several state-of-the-art approaches.**

***Index Terms*—Activity recognition, wearable sensors, deep learning, NMF, stacked auto-encoder.**

I. Introduction

Activity recognition from wearable sensors plays a central role in the development of personalized services in different application domains such as healthcare, pervasive computing, games, and robotics.

Advances in the area of wearable sensing has potential advantages on users everyday life as it enables them to quantify their sleep and exercise patterns [13], monitor personal behaviors [14], track their emotional state [1], [38], [40], [43], and continuously monitor their physiological data in real time [31]. The development of such innovative applications is mainly due to the use of algorithms to infer activities, behaviors and contexts from wearable sensors. Various algorithms have been proposed for human activity recognition from wearable sensors [4], [18], [20], [27], [39]. However, existing algorithms extract hand-crafted features from acceleration data in time or frequency domain to be fed into classification algorithms. The limitations of using hand-crafted features are two fold: 1) they are data dependent, and 2) they could not be generalized for different application domains i.e. suitable features for one application domain might not be suitable for others. To overcome these limitations, researchers started exploring different approaches for automatic feature learning using deep learning algorithms [26], [37], [47]. Deep learning is an emerging area of machine learning that has recently generated significant attention. Deep learning algorithms are capable of learning complex structures and are now key elements in achieving dramatically improved inference performance in a variety of applications such as computer vision [22], natural language processing [41], games [28], and mobile sensing [21].

Very little work has been done on activity recognition using deep learning in supervised way such as deep neural networks [47] or unsupervised way such as auto-encoders [26], [37]. These models are able to automatically learn features from raw acceleration data. However, features are learned using a sliding window for which we need to determine the optimal length, which is data dependent. Moreover, choosing the optimal sliding window length depends on the application domain and could not be generalized to other application domains. For example, in the work of [12], the authors used a sliding window of length 0.28 seconds for the recognition of aggressive and agitated behaviors, whereas in [4], the authors used a sliding window of length 6.7 seconds for the recognition of activities of daily living. This shows how the sliding window length is application dependent.

In this paper, we propose an effective approach for activity recognition from accelerometer data using deep learning. Our approach first applies a dimensionality reduction technique using non-negative matrix factorization to maximize data decorrelation, then it automatically learns features from data using stacked auto-encoders. The recognition is performed using a softmax classifier built on the top hidden layer of the stacked auto-encoder. The deep learning approach allows for in-depth analysis of the underlying data since the new representation implicitly highlights the most informative portions of the analyzed data [37]. The major contributions of this paper can be summarized as follows:

1) Propose an approach to automatically learn suitable features without relying on hand-crafted features and sliding windows using stacked auto-encoders.
2) Combine dimensionality reduction and deep learning in

2375-9259/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDMW.2017.97

one integrated framework to improve activity recognition.

3) Conduct extensive experiments over a variety of publicly available datasets to validate our proposed approach.

The rest of the paper is organized as follows. First, we give an overview of related work in Section 2. Section 3 describes the proposed approach in terms of automatic features extraction, learning and recognition. The results of our experiments on real datasets are presented in Section 4. Finally, Section 5 presents our conclusions and highlights future work directions.

II. Related work

Activity recognition is am important research problem faced by researchers in different domains such as robotics, games, ambient intelligence, and human computer interaction, among others. For example, activity recognition represents a central component in the development of ambient intelligence applications in order to provide appropriate assistance and personalized services for occupants [11].

Much work has been done on activity recognition from wearable sensors in the last decade. Bao et al. [4] proposed a supervised approach for activity recognition using acceleration data collected from 20 subjects wearing five biaxial accelerometers positioned on different parts of the body. The authors extracted four hand-crafted features: mean, energy, frequency-domain entropy, and correlation. Features were computed on 512 sample windows of acceleration data with 256 samples overlapping between consecutive windows. At a sampling frequency of 76.25 Hz, each window represents 6.7 seconds. Features were then fed them into four classifiers such as decision table, K-nearest, neighbor, decision tree, and Naive Bayes. The results obtained showed good performance of decision tree classifier over the others in the recognition of 20 activities. Ravi et al. [39] proposed also a supervised approach for the recognition of eight activities collected from two subjects wearing a triaxial accelerometer near the pelvic region. Four features: mean, standard deviation, energy, and correlation were extracted on 256 sample windows of acceleration data with 128 samples overlapping between consecutive windows. Then, features were used to feed eighteen different classifiers in four different settings for training and testing data. The results obtained showed that Plurality Voting classification performed consistently well across different settings. Parkka et al. [35] proposed an approach based on contextual data such as environmental humidity, environmental temperature, skin temperature, and heart rate collected using different sensors including accelerometers for the recognition of seven activities. Time-domain features such as mean, variance, median, skew, kurtosis, 25% percentile and 75% percentile, and Frequency-domain features such as spectral centroid, spectral spread, estimation of frequency peak, estimation of power of the frequency peak, and signal power in different frequency bands. Time-domain features were extracted using a sliding window. These features were used to feed three classifiers such as automatic decision tree, custom decision tree and artificial neural network. The results obtained showed much better performance of automatic decision tree compared to the other two classifiers. Zhang et al. [49] proposed a framework for activity recognition from wearable sensors using compressed sensing and sparse representation theory. Activity signals were represented as a sparse linear combination of activity signals from all activity classes in the training set. The class membership of the activity signal is determined by solving a ℓ_1 minimization problem. Many features such as mean, median, variance, standard deviation, first order derivative, second order derivative, kurtosis, skewness, zero crossing rate, mean crossing rate, energy, correlation, and velocity, among others were extracted using a sliding window of size 4 seconds with 50% overlap. The results obtained showed that the classification via sparse representation performed better than conventional classifiers such as SVM, KNN, and Naive Bayes. Altun et al. [2] proposed Bayesian decision making based model for the recognition of 19 activities of daily living performed by eight subjects using accelerometers placed on five different locations of the subject's body. Features such as mean, variance, skewness, kurtosis, autocorrelation, and Discrete Fourier transform were extracted using a sliding window of size 5 seconds.

With the tremendous growth of smart phones with accelerometers mounted in, several approaches were proposed for the activity recognition using smart phones. Kose et al. [29] proposed a real-time approach using a smart phone's acceleration data collected from five subjects for the recognition of four activities. Features such as average, minimum, maximum, and standard deviation were extracted using a sliding window. Features were then fed into a clustered KNN classifier. The results obtained showed better performance of the clustered KNN compared to the Naive Bayes classifier. Kwapisz et al. [19] collected data from 29 users using smart phones worn on the front pant pocket. Features such as average acceleration, average absolute difference, average resultant acceleration, time between peaks, Binned distribution and standard deviation were extracted using a sliding window of size 10 seconds to feed three classifiers: decision tree, logistic regression, and three-layer neural network. The later classifier showed best performance. Bayat et al. [5] proposed an activity recognition system from smart phones. The authors designed a low-pass filter to isolate the component of gravity acceleration form the body acceleration in raw data. Features such as mean, elapsed time between consecutive local peaks, average of peak frequency (APF), variance of APF, root mean square (RMS), standard deviation, minimum, maximum, and correlation were extracted using a sliding window of size 128 with 50% overlap. These features were then fed into different classifiers such as Multilayer Perceptron, SVM, Random Forest, Logistic Model Trees, Simple Logistic, and Logit Boost. The results showed that the combination of different classifiers together such as SVM, Multilayer Perceptron, and Logit Boost perform better than the other classifiers. Other approaches on activity recognition using smart phones are discussed in the survey of [42].

Deep learning approaches have been used to overcome the problem of hand-crafted features in activity recognition. However, very little work has been done on activity recognition from acceleration data using deep learning. Zhang et al. [47] proposed deep neural networks for activity recognition from wearable sensors. Deep neural networks were able to automatically learn features from data using a sliding window of size 1 second with 50% overlap. The results obtained showed that deep neural networks outperformed hand-crafted based approaches. Plötz et al. [37] used Deep Belief Networks combined with PCA to learn features for activity recognition in ubiquitous environments. A data representation technique based on the empirical cumulative distribution functions was used to derive a representation of the input data, which is independent of the absolute ranges but preserves structural information of sensor data. A window size of 64 samples was used with 50% of data overlapping. Li et al. [26] proposed an unsupervised feature learning approach for activity recognition from acceleration data using deep learning. Auto-encoders were used to learn features automatically in an unsupervised way. Three methods were used for this purpose such as sparse auto-encoders, denoising auto-encoders and PCA. the results obtained showed that the sparse auto-encoder achieved better results than the other two techniques. However, due the to unsupervised nature of their approach, the authors did not take into account the label information contained in the data, which could significantly improve the activity recognition performance. Wang [45] proposed an approach for recognizing human activities using continuous auto-encoder by adding Gaussian random units into the sigmoid activation function to extract the features of nonlinear data. Hongqing et al. [16] proposed a recurrent neural network (RNN) based model for activity recognition from smart phones. The authors used hand-crafted features such as sensor ID, time of the day, and activity length to train the RNN model. However, the approach extracts first time and frequency domain features manually from the original data using a 5 seconds window, which makes the approach difficult to be generalized to other datasets. Ordez et al. [32] proposed a model that combines deep convolutional networks and long short term memory neural network for multimodal wearable activity recognition. Their model is very resource-demanding and time consuming given the combination of two complex neural network models.

The approaches described above suffered from one of the following limitations: 1) they are based on hand-crafted features extracted manually by the authors. These features are data dependent and are different for each application and context, which makes it hard to find the suitable features that could be generalized across different application domains; 2) the features are extracted using sliding windows, which are data dependent and are designed by the authors. This makes it difficult to find the optimal window size that could work for different applications. These points motivate us to propose a new principled approach activity recognition from accelerometer data that addresses the limitations of the existing approaches. Our approach combines dimensionality reduction method and deep learning for accurate activity representation and recognition. The discrimination power of non-negative matrix factorization, and the performance of deep learning algorithms compared to traditional data mining algorithms will help strengthen our approach and make it effective and generalizable compared to the existing approaches.

III. Proposed Approach

In this section, we describe our approach for human activity recognition in terms of non-negative matrix factorization, automatic feature extraction and classification. Figure 1 shows an overview of the different steps of our approach. The details of each segment in Figure 1 are presented in the following sections.

A. NMF Based Data Representation

Non-negative Matrix Factorization (NMF) is a matrix factorization algorithm that finds the positive factorization of a given positive matrix [23], [24]. In NMF, each axis captures the base information of a particular behavior class, and each behavior is represented as an additive combination of the base information. The class membership of each behavior can be easily determined by finding the base posture (the axis) with which the behavior has the largest projection value. Therefore, the potential of using NMF lies in the discriminative power between the behaviors when projected into the new space. NMF has been successfully applied in different situations such as parts-based representation in human brain [34], learning parts of objects like human faces [33], face recognition [25] and document clustering [46] among others.

Formally, given a data matrix $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{m \times n}$, NMF consists in factorizing the matrix $\mathbf{X}$ into the non-negative matrix $\mathbf{U} = [u_{ij}] \in \mathbb{R}^{m \times k}$ and the non-negative matrix $\mathbf{V} = [v_{ij}] \in \mathbb{R}^{n \times k}$ as follows:

$$\mathbf{X} \approx \mathbf{U}\mathbf{V}^T, \tag{1}$$

by minimizing the following objective function Φ:

$$\Phi = \frac{1}{2} \parallel \mathbf{X} - \mathbf{U}\mathbf{V}^T \parallel \tag{2}$$

where $\parallel . \parallel$ denotes the squared sum of all in the matrix (please see section IV-G3 on how to select the rank k). Here, the objective function Φ, which represents the squared Euclidean distance, seeks to minimize the error of the reconstruction of the original matrix $\mathbf{X}$ by the product $\mathbf{UV}$. The objective function Φ can be rewritten as follows:

$$\begin{aligned}\Phi &= \frac{1}{2}tr((\mathbf{X} - \mathbf{U}\mathbf{V}^T)(\mathbf{X} - \mathbf{U}\mathbf{V}^T)^T) \\ &= \frac{1}{2}tr(\mathbf{X}\mathbf{X}^T - 2\mathbf{X}\mathbf{V}\mathbf{U}^T + \mathbf{U}\mathbf{V}^T\mathbf{V}\mathbf{U}^T) \\ &= \frac{1}{2}(tr(\mathbf{X}\mathbf{X}^T) - 2tr(\mathbf{X}\mathbf{V}\mathbf{U}^T) + tr(\mathbf{U}\mathbf{V}^T\mathbf{V}\mathbf{U}^T))\end{aligned} \tag{3}$$

here the matrix property $tr(\mathbf{UV}) = tr(\mathbf{VU})$ is used in the derivation steps. Lee & Seung [24] presented an iterative

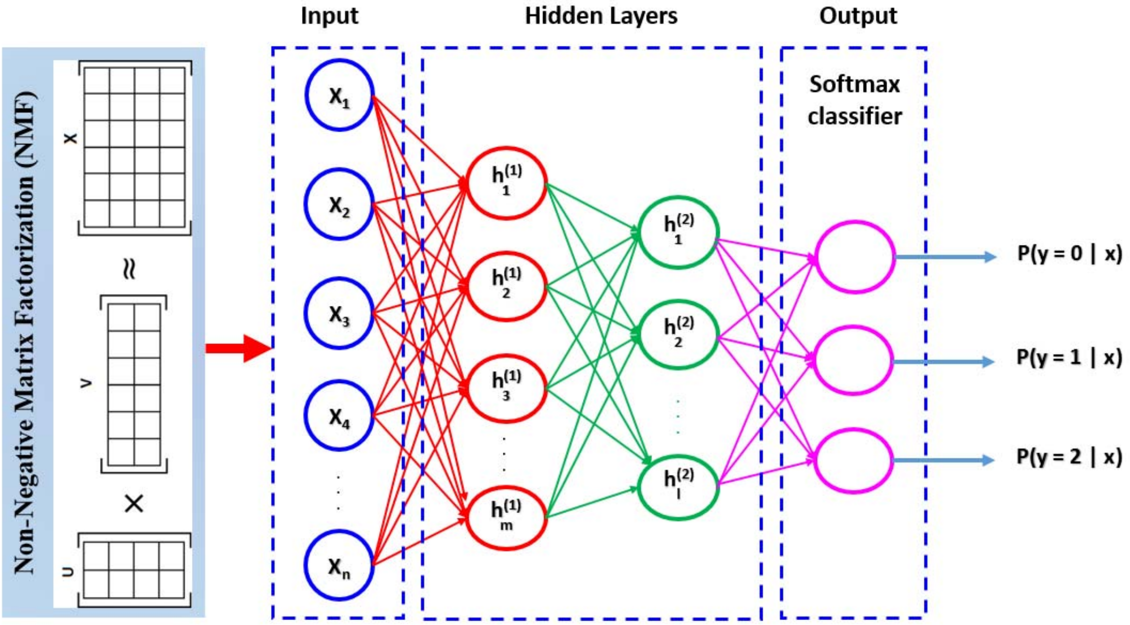

Fig. 1. Overview of the different steps of our approach.

update algorithm to find a local minimum of the objective function Φ as follows:

$$u_{ij}^{t+1} = u_{ij}^{t} \frac{(\mathbf{XV})_{ij}}{(\mathbf{UV}^{T}\mathbf{V})_{ij}} \tag{4}$$

$$v_{ij}^{t+1} = v_{ij}^{t} \frac{(\mathbf{X}^{T}\mathbf{U})_{ij}}{(\mathbf{VU}^{T}\mathbf{U})_{ij}} \tag{5}$$

Lee & Seung [24] proved that the convergence of the iterations is guaranteed, however, the solution to minimizing the objective function Φ is not unique. If $\mathbf{U}$ and $\mathbf{V}$ are the solutions to Φ, then, $\mathbf{UH}$ and $\mathbf{VH}^{-1}$ will also form a solution for any positive diagonal matrix $\mathbf{H}$. To this end, a normalization is needed to make the solution unique as follows:

$$u_{ij} = \frac{u_{ij}}{\sqrt{\sum_i u_{ij}^2}} \tag{6}$$

$$v_{ij} = v_{ij} \sqrt{\sum_i u_{ij}^2} \tag{7}$$

Therefore, each data vector $\mathbf{x}_i$ is approximated by a linear combination of the columns of $\mathbf{U}$, weighted by the components of $\mathbf{V}$. The non-negative constraints on $\mathbf{U}$ and $\mathbf{V}$ allow additive combinations among different basis. Unlike SVD, no substraction can occur in NMF. This is the most significant difference between NMF and other matrix factorization algorithms such as SVD, PCA, and vector quantization (VQ) [8]. For instance, in VQ, each column of $\mathbf{V}$ is constrained to be a unary vector, i.e. one element equal to unity and the remaining elements equal to zero. In PCA the columns of $\mathbf{U}$ are constrained to be orthonormal and the rows of $\mathbf{V}$ to be orthogonal to each other, which is considered as relaxation of the unary property in VQ [23]. In contrast, NMF does not allow negative entries in both matrices $\mathbf{U}$ and $\mathbf{V}$. The non-negativity property of NMF allows the combination of multiple base information of basic tasks to represent the human activity.

B. Automatic Feature Extraction Using stacked Auto-encoders

Here we first briefly define the traditional auto-encoder, then we introduce the stacked auto-encoders. We use the same formulation used in [44]. An auto-encoder is a neural network with a single hidden layer. An auto-encoder is composed of two main steps:

- **Encoder**: is a deterministic mapping that transforms an input vector $\mathbf{x}$ into a hidden representation $\mathbf{y} = f_{\theta}(\mathbf{x})$ as follows:

  $$f_{\theta}(\mathbf{x}) = s(\mathbf{Wx} + \mathbf{b}) \tag{8}$$

  $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{b}\}$ are the parameter set, where $\mathbf{W}$ is $d' \times d$ weight matrix and **b** is an offset vector of dimensionality d. s(.) is a sigmoid function (the activation function).
- **Decoder**: the hidden representation $\mathbf{y}$ obtained in the encoding step is then mapped back to a reconstructed d-dimensional vector $\mathbf{z}$ in input space, $\mathbf{z} = g_{\theta'}(\mathbf{y})$. The

mapping $g_{\theta'}$ is called the **decoder** and can be written as follows:

$$g_{\theta'}(\mathbf{y}) = s(\mathbf{W}'\mathbf{y} + \mathbf{b}') \tag{9}$$

$\boldsymbol{\theta}' = \{\mathbf{W}', \mathbf{b}'\}$ are the parameter set.

Note that **z** is not an exact reconstruction of the input **x**. It can be interpreted as the parameters of a distribution p(X—Z=**z**) that might generate **x** with high probability, which yields a reconstruction error to be minimized:

$$L(\mathbf{x}, \mathbf{z}) \propto -logp(\mathbf{x}|\mathbf{z}) \tag{10}$$

A stacked auto-encoder is a deep network consisting of multiple layers of auto-encoders in which the outputs of each layer is wired to the inputs of the successive layer as shown in Figure 1. Formally, consider a stacked auto-encoder with n layers. Let $\mathbf{W}^{(k,1)}$, $\mathbf{W}^{(k,2)}$, $\mathbf{b}^{(k,1)}$, $\mathbf{b}^{(k,2)}$ denote the parameters $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$, $\mathbf{b}^{(1)}$, $\mathbf{b}^{(2)}$ for k^{th} auto-encoder. Then the encoding step for the stacked auto-encoder is given by running the encoding step of each layer in forward order as follows:

$$a^{(l)} = f(z^{(l)}) \tag{11}$$

$$z^{(l+1)} = \mathbf{W}^{(l,1)}a^{(l)} + \mathbf{b}^{(l,1)} \tag{12}$$

The idea of our work is to build a stacked auto-encoder to automatically extract features from different activities in an unsupervised way. We take advantage of activity labels available in datasets to perform a supervised learning by connecting the top hidden layers of stacked auto-encoders to activity labels. Therefore, a classifier is built on the top of the last hidden layer of the encoding step. In our work, we use a softmax classifier to classify behaviors as shown in Figure 1. Formally, the softmax classifier can be defined as follows:

$$Softmax(z)_i = \frac{exp(z_i)}{\sum_{l=1}^{C} exp(z_l)} \tag{13}$$

where z_i represents the i^{th} element of the input to softmax, which corresponds to class i, and C is the number of classes.

IV. Validation

We evaluate the performance of our approach on six publicly available real human behavior datasets described below.

A. Opportunity Dataset

The Opportunity dataset [9] relates to a home environment (kitchen) and the analysis of activities of daily living using multiple worn and embedded sensors. Four subjects participated in data collection on different days by wearing multiple accelerometers on different locations of the body such as right arm, left arm, right wrist, left wrist, hip and back. Our analysis was based on the sensor data recorded by the accelerometer attached to the right arm, left arm of the subject and both arms together. We considered 5 low-level activities of interest. We used ADL1, ADL2, ADL3, and Drill for training, and ADL4 and ADL5 for testing as recommended by the authors. The acceleration data was sampled with 64Hz sampling rate.

B. USC Dataset

The USC dataset [50] relates to 12 human basic activities collected by 14 subjects by wearing an accelerometer sensor on the subjects' front right hip. We choose leave one out cross validation method for training and testing (i.e. we choose one subject for testing and all the remaining subjects for training, and we redo this process for every subject). The acceleration data was sampled with 64Hz sampling rate.

C. Sports and Daily Activities Dataset

In this dataset [2], 19 daily living and sports activities were collected by 8 subjects using accelerometer sensors placed on five different places on the subject's body. Each activity was performed for 5 minutes. A sliding window of size 5 seconds was used to divide each activity into segments for feature extraction. A total of 60 segments were generated for each activity for each subject yielding (60 × 8 = 480) segments for each activity in the entire dataset. The acceleration data was sampled with 25Hz sampling rate.

D. Berkeley MHAD Dataset

The Berkeley Multimodal Human Action Database (MHAD) [30] contains 11 actions performed by 12 subjects. Six three-axis wireless accelerometers were placed on different locations of the subject's body to measure movements at the wrists, ankles, and hips. Each accelerometer sequence data was partitioned using a sliding window of size 15. All the subjects performed 5 repetitions of each action, yielding about 660 action sequences which correspond to about 82 minutes of total recording time. The acceleration data was sampled with 30Hz sampling rate.

E. Human Motion Dataset

The Human Motion Dataset (HMD) [7] is composed of the recordings of 14 simple ADL (brush teeth, climb stairs, comb hair, descend stairs, drink glass, eat meat, eat soup, getup bed, liedown bed, pour water, sitdown chair, standup chair, use telephone, walk) performed by a total of 16 subjects. The data are collected by a single tri-axial accelerometer attached to the right-wrist of the subject. The acceleration data was sampled with 32Hz sampling rate.

F. UTD-MHAD Dataset

The UTD-MHAD dataset [10] is composed of the recordings of 27 simple gestures and sports motions ((1) right arm swipe to the left, (2) right arm swipe to the right, (3) right hand wave, (4) two hand front clap, (5) right arm throw, (6) cross arms in the chest, (7) basketball shoot, (8) right hand draw x, (9) right hand draw circle (clockwise), (10) right hand draw circle (counter clockwise), (11) draw triangle, (12) bowling (right hand), (13) front boxing, (14) baseball swing from right, (15) tennis right hand forehand swing, (16) arm curl (two arms), (17) tennis serve, (18) two hand push, (19) right hand knock on door, (20) right hand catch an object, (21) right hand pick up and throw, (22) jogging in place, (23) walking in place, (24) sit to stand, (25) stand to sit, (26) forward lunge (left foot forward), (27) squat (two arms stretch out)) performed by a total of 8 subjects. The data are collected by a 9-axis MEMS sensor which captures 3-axis acceleration, 3-axis angular velocity and 3-axis magnetic strength. The acceleration data was sampled with 50Hz sampling rate.

G. Experimental setup

We first evaluate the performance of our proposed approach using the six previously described datasets. Then, we compare our results to the state-of-the-art methods to demonstrate the superiority and effectiveness of our proposed approach. In our experiments, we used different measures such as precision, recall and F-Measure to evaluate the performance of our approach. We experimentally determined the optimal rank of the NMF method $k = 2$ that achieved the best classification results (please see section IV-G3 for more details). Determining the optimal factorization rank automatically will be considered in our future work. For the stacked auto-encoder model, the inputs were the results of the NMF factorization method. The training set is divided into mini-batches each having about 100 frames. We used the Matlab Neural Network toolbox to implement our model. The training is performed with a learning rate of 0.0005. We experimentally determined that 2 is the optimal number of hidden layers for the stacked auto-encoder. We used 100 units in the first layer and 50 units in the second layer. For the softmax classifier, we used 1000 epochs in the training using the gradient descent with a learning rate of 0.0001.

1) Leave One Out Cross Validation: In this experiment, we used all behavior instances from one (1) participant for testing and the behavior instances of the remaining participants for testing. We performed the experiments NP times (where NP is the number of participants in each dataset), excluding one participant at each time. The benefit of such setup is twofold. First, it allows detecting problematic participants and analyzing the sources of some of the classification errors caused by these participants. A problematic participant means his/her behaviors were performed differently compared to other participants. Second, it allows testing the inter-participant generalization of the approach, which constitutes a good indicator about the practicability of our approach. Table I shows the recognition results obtained for each dataset.

TABLE I
RECOGNITION RESULTS OBTAINED FOR EACH DATASET.

Dataset	Precision	Recall	F-measure
$Sport - ADL$	0.9868	0.9901	0.9885
$Opportunity$	0.9994	0.9994	0.9994
$USC - dataset$	0.996	0.997	0.9965
$Berkeley - MHAD$	0.9903	0.992	0.9911
$Human - motion$	0.9965	0.9987	0.9976
$UTD - MHAD$	0.9259	0.9275	0.9267

The results obtained in all the datasets are very interesting except for the UTD-MHAD dataset. The good results obtained can be explained by the fact that, the NMF method decorrelates perfectly the data, which greatly helps discriminating between the different activities and learning the representation of each activity using the stacked auto-encoder. For the UTD-MHAD dataset, the results were not as good as those obtained in the other datasets because 1) of the high number of activities in this dataset, and (2) the similarity between some activities such as jogging in place and walking in place, and between sit to stand and stand to sit, which leads the model to make confusions between these activities. Figure 2 shows the confusion matrix (in terms of accuracy) for each activity in the UTD-MHAD dataset.

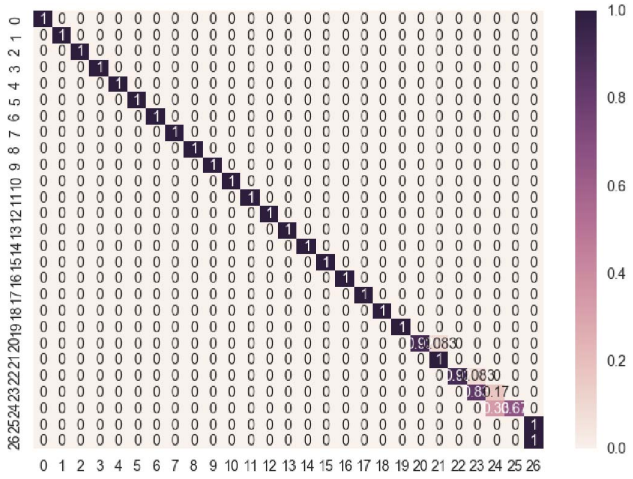

Fig. 2. Confusion matrix obtained for the UTD-MHAD dataset.

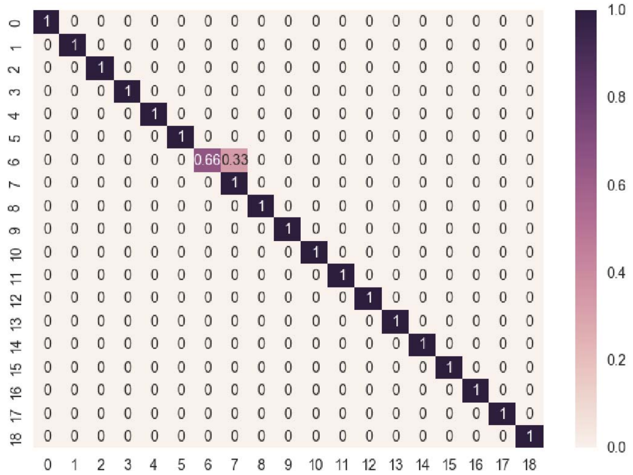

Fig. 3. Confusion matrix obtained for the Sport dataset.

Similarly, in the Sport and daily living activities dataset, the similarity between the activities standing in an elevator still (A7) and moving around in an elevator (A8) leads the model to create confusion and to incorrectly classify activity A7 as activity A8 and inversely. However, the model correctly classifies all the remaining activities in this dataset. Figure 3 shows the confusion matrix obtained for the Sport and daily living activities dataset.

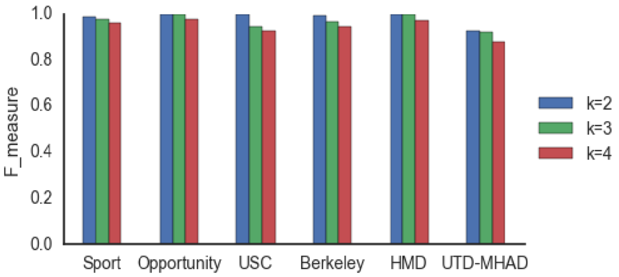

Fig. 4. Recognition accuracy using different values of NMF rank.

The variability observed in the ways participants performed the different activities constitutes a good validation setting for our approach. This is demonstrated by the promising results obtained using datasets with large number of activities such as Sport and UTD-MHAD datasets.

2) NMF vs PCA and SVD: In this section we report the results obtained by comparing NMF with well known matrix factorization techniques such as PCA and SVD. To do so, we replaced NMF in our model by PCA first, and then SVD. We used the leave one out method experimental setting to evaluate the results. Table II shows a comparison of the activity recognition results obtained using PCA and SVD.

TABLE II
RECOGNITION RESULTS OBTAINED FOR EACH DATASET.

	PCA			SVD		
Dataset	Precision	Recall	F-measure	Precision	Recall	F-measure
$Sport - ADL$	0.6493	0.6328	0.6410	0.5515	0.5661	0.5586
$Opportunity$	0.5404	0.5404	0.5404	0.3969	0.5839	0.4726
$USC - dataset$	0.6724	0.6211	0.6457	0.8434	0.8073	0.8250
$Berkeley - MHAD$	0.6542	0.6743	0.6629	0.82	0.8580	0.8382
$Human - motion$	0.2908	0.2576	0.2732	0.7213	0.7747	0.7470
$UTD - MHAD$	0.3282	0.3351	0.3316	0.5034	0.5240	0.5135

As shown in Table II, SVD performs better than PCA in all datasets except for Sport-ADL and Opportunity datasets, where PCA performs relatively better. However, NMF performs better than Both methods as shown in Table I. In fact, the potential of the NMF method lies in the ability to identify potentially relevant features that represent local and global characteristics of each activity. In addition, with the NMF method, activities are represented as additive combination of the base movements, which matches the reality of doing daily living activities.

3) NMF rank selection: A critical parameter in NMF is the factorization rank. Choosing the optimal rank for initializing NMF is crucial for the performance of the NMF algorithm. A common way of choosing the rank is to try different values, compute some quality measure of the results, and choose the best value accordingly. In our work, we used the F-Measure as a quality measure. Figure 4 shows how the F-Measure varies by varying the rank of the NMF technique using all datasets.

We observe from Figure 4 that the F-Measure is high when the rank of NMF is small (rank = 2). The accuracy decreases by increasing the value of the NMF rank, which means that the discrimination ability of NMF is higher in low dimensional space. However, the discrimination ability between the different behaviors decreases when the dimension of the space increases. Besides, performing a NMF factorization with high rank values is time consuming and computationally ineffective. It has been shown that low values of the NMF rank achieved better performance compared to high values [6], [17]. This is also the case in our approach where rank 2 achieves the best performance. Interestingly though, when the rank of NMF increases the F-Measure decreases. This is an important observation, which means that when the rank of NMF is greater than 2, the projection into the new space does not change the discrimination ability of NMF. This suggests the need for an automatic method that takes into account both the accuracy and the computational complexity in selecting the optimal NMF rank.

H. Comparison With State-of-the-Art Methods

We compared our approach with several existing approaches in the literature. These approaches include frequently used classifiers such as Decision tree, Random forests, Bayesian nets, Naive Bayes, Support vector machines, K-nearest neighbor, K-means clustering and Perceptron neural networks. These approaches use hand-crafted features such as mean, variance, correlation, entropy, energy, kurtosis and skewness to construct a training and test datasets. However, in the deep learning based approaches such as Deep belief networks, Convolutional neural networks, stacked auto-encoders and Recurrent neural networks, features are automatically extracted from data in an unsupervised way. The comparison results are presented in Table III for each dataset.

We compared our approach with the approach of [48] by constructing a deep belief network (DBN). The DBN model was pre-trained using stochastic gradient decent with a mini-batch at a time. We ran 100 epochs for the Gaussian-binary RBM at learning rate 0.001 and ran 50 epochs for the binary-binary RBMs at learning rate 0.1. The constructed DBN model had 500 units in the first layer, 500 in the second layer and 2000 in the third layer. Given the complexity level of this model, we reported the results only for three datasets where the model was able to produce results as shown in Table III. For the remaining datasets, the model was not able to produce any results and crashes most of the time because of its complexity particularly for the third layer with 2000 units, which makes the model very resource-demanding.

We have also compared our approach with two different deep neural networks such as recurrent neural networks (RNN) and convolutional neural networks (CNN). We used Tensorflow google machine learning library to implement these models. The results obtained are shown in Table III.

The results obtained show clearly the ability of our approach to discriminate between the different activities and its superiority compared to the other approaches. As shown in Table III, the only method that achieves better results compared to our approach is the method of Ravi et al. [43] using decision tree classifier with an F-measure greater than 0.98 in the UTD-MHAD dataset. This method achieves also good results in the

TABLE III

COMPARISON OF THE RECOGNITION ACCURACY RESULTS OBTAINED FROM THE CONVENTIONAL CLASSIFIERS AND OUR APPROACH.

Approach	**Classifier**	**Results**					
		Berkeley Dataset	USC Dataset	HMD Dataset	Opportunity Dataset	UTD MHAD Dataset	Sport Dataset
		F-Measure	F-Measure	F-Measure	F-Measure	F-Measure	F-Measure
Zhang et al. [48]	Deep belief networks (DBN)	Na	0.917	Na	0.823	Na	0.906
Baseline	Recurrent neural networks (RNN)	0.3832	0.8602	0.4345	0.5816	0.2061	0.4326
Baseline	Convolutional neural networks (CNN)	0.6844	0.9233	0.5344	0.912	0.5509	0.7020
Ermes et al. [15]	Decision tree	0.6242	0.7384	0.3970	0.8371	0.3328	0.4864
	Neural network	0.6763	0.7095	0.4454	0.7964	0.3375	0.5383
Pirttikangas et al. [36]	K nearest neighbors	0.6705	0.8257	0.4107	0.9049	0.3384	0.4031
	Multilayer perceptron	0.6585	0.6279	0.4444	0.7681	0.3455	0.5626
Bao et al. [4]	Decision Tables	0.8269	0.6018	0.3699	0.3831	0.3093	0.2712
	Decision Trees	0.7576	0.7368	0.4281	0.7922	0.3537	0.4462
	K nearest neighbors	0.545	0.8099	0.4108	0.7733	0.3661	0.3589
	Naive Bayes	0.4764	0.3838	0.4689	0.8399	0.4411	0.4603
Ravi et al. [43]	Decision Tables	0.7968	0.7357	0.4149	0.8702	0.7968	0.2170
	Decision Trees	0.9844	0.9424	0.4619	0.9784	**0.9844**	0.4470
	K nearest neighbors	0.9844	0.9289	0.5008	0.9704	0.6210	0.4559
	Naive Bayes	0.8605	0.5371	0.4604	0.9704	0.8605	0.6272
	SVM	0.9889	0.5115	0.4492	0.2490	0.1826	0.6248
Atallah et al. [3]	K nearest neighbors (K=5)	0.9350	0.8315	0.4813	0.968	0.5441	0.4949
	K nearest neighbors (K=7)	0.9086	0.8258	0.4813	0.9574	0.5619	0.4786
	Bayesian classifier	0.8791	0.6668	0.4547	0.9413	0.5323	0.5334
Our approach	NMF+SAE	**0.9911**	**0.9965**	**0.9976**	**0.9994**	0.9267	**0.9885**

Opportunity, USC and Berkeley datasets with an F-measure of 0.9784, 0.9424 and 0.9844 respectively.

An important observation lies in the method of Pirttikangas et al. that employed also a K nearest neighbors classifier, but the results were very low compared to the method of Ravi et al. For example, in the Opportunity dataset, Pirttikangas et al's method using K nearest neighbors classifier achieves an F-measure of 0.9049, whereas the method of Ravi et al. achieves an F-measure of 0.9704 using the same classifier. This can be explained by the set of features employed in each method such as the energy and correlation between X and Z and Y and Z axis features that have not been used in the Pirttikangas et al's method.

For the deep neural network methods, only the DBN model achieves good results in three datasets such as USC, Opportunity and Sport with an F-measure of 0.917, 0.823 and 0.906 respectively. The RNN and CNN models perform poorly in all the datasets except for the USC dataset where they achieve good results. The USC dataset is characterized by a huge number of instances for each activity, which explains the good performance of deep neural network methods as well as traditional approaches, compared to the HMD and UTD-MHAD datasets where activities have small number of instances.

I. Effect of the number of hidden layers

The number of hidden layers plays an important role in learning features representations. However, adding more hidden layers will significantly increase model complexity and execution time. In this section, we evaluate the performance of our model by increasing the number of hidden layers. We show results only for the USC dataset. Table IV shows the results obtained in terms of precision, recall and F-measure by varying the number of hidden layers.

As shown in Table IV, our model performs better with 2 hidden layers, then the performance decreases when the number of hidden layers equals to 3. After that, the performance increases significantly when the number of hidden layers is 4, 5 and 6 and stabilizes after that. Interestingly though, increasing the number of hidden may increase performance, however, the model becomes more complex and resource-demanding. In our model, we choose 2 as the optimal number of hidden layers that makes a tradeoff between the model performance and complexity.

TABLE IV

PERFORMANCE OF OUR MODEL WITH DIFFERENT NUMBER OF HIDDEN LAYERS IN THE USC DATASET.

Number of hidden layers	Precision	Recall	F-measure
1	0.8304	0.869	0.8492
2	0.996	0.997	0.9965
3	0.9056	0.9192	0.9123
4	0.9464	0.9601	0.9532
5	0.9663	0.9749	0.9705
6	0.9782	0.9836	0.9808
7	0.9881	0.9905	0.9892
8	0.9911	0.9929	0.9919
9	0.9926	0.994	0.9932
10	0.9911	0.9929	0.9919

V. CONCLUSION

In this paper we have studied the problem of human activity recognition using deep learning. We have proposed an effective approach based on non-negative matrix factorization and stacked auto-encoders. Our approach applies first a matrix factorization in order to project data into a new reduced space to find a best activity representation and to increase the discrimination ability of our approach. Then, features were automatically extracted from the projected data using stacked auto-encoders. For classification, we build a softmax classifier on the top hidden layer of the stacked auto-encoder.

We have illustrated the effectiveness and suitability of our approach through extensive experiments on multiple publicly available real human activity datasets. The experimental results show the suitability of our approach in representing activities and distinguishing between them. In addition, we have also illustrated how our approach outperformed several of the state-of-the-art methods. We empirically demonstrated the effectiveness of the non-negative matrix factorization over the PCA and SVM factorization methods when combined with stacked auto-encoders.

For future work, we will evaluate the suitability of our approach for cross datasets transfer learning. This will greatly helpful 1) to address the automatic labeling challenge of new datasets, 2) to perform activity recognition on large-scale data, and 3) to ease the deployment of activity recognizers on cloud infrastructures.

References

[1] T. Ahram, W. Karwowski, D. Schmorrow, L. Paletta, N. Pittino, M. Schwarz, V. Wagner, and K. Kallus, "6th international conference on applied human factors and ergonomics (ahfe 2015) and the affiliated conferences, ahfe 2015 human factors analysis using wearable sensors in the context of cognitive and emotional arousal," *Procedia Manufacturing*, vol. 3, pp. 3782 – 3787, 2015.

[2] K. Altun, B. Barshan, and O. Tunçel, "Comparative study on classifying human activities with miniature inertial and magnetic sensors," *Pattern Recogn.*, vol. 43, no. 10, pp. 3605–3620, 2010.

[3] L. Atallah, B. Lo, R. King, and G. Z. Yang, "Sensor placement for activity detection using wearable accelerometers," in *Body Sensor Networks (BSN), 2010 International Conference on*, 2010, pp. 24–29.

[4] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *Pervasive Computing, Second International Conference, PERVASIVE 2004, Vienna, Austria, Proceedings*, 2004, pp. 1–17.

[5] A. Bayat, M. Pomplun, and D. A. Tran, "A study on human activity recognition using accelerometer data from smartphones," *Procedia Computer Science*, vol. 34, pp. 450–457, 2014.

[6] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the National Academy of Sciences*, vol. 101, no. 12, pp. 4164–4169, 2004.

[7] B. Bruno, F. Mastrogiovanni, A. Sgorbissa, T. Vernazza, and R. Zaccaria, "Analysis of human behavior recognition algorithms based on acceleration data," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, 2013, pp. 1602–1607.

[8] D. Cai, X. He, X. Wu, and J. Han, "Non-negative matrix factorization on manifold," in *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on*, 2008, pp. 63–72.

[9] R. Chavarriaga, H. Sagha, A. Calatroni, S. T. Digumarti, G. Trster, J. del R. Milln, and D. Roggen, "The opportunity challenge: A benchmark database for on-body sensor-based activity recognition," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 2033 – 2042, 2013.

[10] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *2015 IEEE International Conference on Image Processing (ICIP)*, Sept 2015, pp. 168–172.

[11] B. Chikhaoui, S. Wang, and H. Pigot, "ADR-SPLDA: activity discovery and recognition by combining sequential patterns and latent dirichlet allocation," *Pervasive and Mobile Computing*, vol. 8, no. 6, pp. 845–862, 2012.

[12] B. Chikhaoui, B. Ye, and A. Mihailidis, "Feature-level combination of skeleton joints and body parts for accurate aggressive and agitated behavior recognition," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–20, 2016.

[13] S. Consolvo, D. W. McDonald, T. Toscos, M. Y. Chen, J. Froehlich, B. Harrison, P. Klasnja, A. LaMarca, L. LeGrand, R. Libby, I. Smith, and J. A. Landay, "Activity sensing in the wild: A field trial of ubifit garden," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2008, pp. 1797–1806.

[14] C. Emil, R. Carlos, and D. Pronabesh, "Patient-centered activity monitoring in the self-management of chronic health conditions," *BMC Medicine*, vol. 13, no. 1, pp. 1–6, 2015.

[15] M. Ermes, J. Parkka, J. Mantyjarvi, and I. Korhonen, "Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 12, no. 1, pp. 20–26, 2008.

[16] H. Fang, H. Si, and L. Chen, "Recurrent neural network for human activity recognition in smart home," in *Proceedings of 2013 Chinese Intelligent Automation Conference: Intelligent Automation*, Z. Sun and Z. Deng, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 341–348.

[17] B. Kanagal and V. Sindhwani, "Rank selection in low-rank matrix approximations: A study of cross-validation for nmfs," *Reconstruction*, vol. 1, pp. 1–10, 2010.

[18] D. M. Karantonis, M. R. Narayanan, M. Mathie, N. H. Lovel, and B. G. Celler, "Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 10, no. 1, pp. 156–167, 2006.

[19] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," in *SensorKDD' 2010*, 2010, pp. 74–82.

[20] ——, "Activity recognition using cell phone accelerometers," *SIGKDD Explor. Newsl.*, vol. 12, no. 2, pp. 74–82, 2011.

[21] N. D. Lane and P. Georgiev, "Can deep learning revolutionize mobile sensing?" in *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, 2015, pp. 117–122.

[22] B. Y. LeCun Yann and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[23] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[24] ——, "Algorithms for non-negative matrix factorization," in *In NIPS*. MIT Press, 2000, pp. 556–562.

[25] S. Li, X. Hou, H. Zhang, and Q. Cheng, "Learning spatially localized, parts-based representation," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, 2001, pp. I–207–I–212 vol.1.

[26] Y. Li, D. Shi, B. Ding, and D. Liu, *Unsupervised Feature Learning for Human Activity Recognition Using Smartphone Sensors*. Springer International Publishing, 2014, pp. 99–107.

[27] J. Liu, L. Zhong, J. Wickramasuriya, and V. Vasudevan, "uwave: Accelerometer-based personalized gesture recognition and its applications," *Pervasive and Mobile Computing*, vol. 5, no. 6, pp. 657 – 675, 2009.

[28] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. A. Riedmiller, "Playing atari with deep reinforcement learning," *CoRR*, vol. abs/1312.5602, 2013.

[29] C. E. Mustafa Kose, Ozlem Durmaz Incel, "Online human activity recognition on smart phones," in *2nd International Workshop on Mobile Sensing*, 2012, pp. 11–15.

[30] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, 2013, pp. 53–60.

[31] N. Oliver and F. Flores-Mangas, "Healthgear: a real-time wearable system for monitoring and analyzing physiological signals," in *International Workshop on Wearable and Implantable Body Sensor Networks (BSN'06)*, 2006, pp. 4 pp.–64.

[32] F. J. Ordez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, 2016.

[33] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[34] S. E. Palmer, "Hierarchical structure in perceptual representation," *Cognitive Psychology*, vol. 9, no. 4, pp. 441 – 474, 1977.

[35] J. Parkka, M. Ermes, P. Korpipaa, J. Mantyjarvi, J. Peltola, and I. Korhonen, "Activity classification using realistic data from wearable sensors," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 1, pp. 119–128, 2006.

[36] S. Pirttikangas, K. Fujinami, and T. Nakajima, "Feature selection and activity recognition from wearable sensors," in *Ubiquitous Computing Systems, Third International Symposium, UCS 2006, Seoul, Korea, October 11-13, 2006, Proceedings*, 2006, pp. 516–527.

[37] T. Plötz, N. Y. Hammerla, and P. Olivier, “Feature learning for activity recognition in ubiquitous computing,” in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, ser. IJCAI’11, 2011, pp. 1729–1734.

[38] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas, “Emotionsense: A mobile phones based adaptive platform for experimental social psychology research,” in *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, 2010, pp. 281–290.

[39] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, “Activity recognition from accelerometer data,” in *AAAI*, vol. 5, 2005, pp. 1541–1546.

[40] A. Sano and R. W. Picard, “Stress recognition using wearable sensors and mobile phones,” in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, 2013, pp. 671–676.

[41] R. Sarikaya, G. E. Hinton, and A. Deoras, “Application of deep belief networks for natural language understanding,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 4, pp. 778–784, 2014.

[42] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. Havinga, “A survey of online activity recognition using mobile phones,” *Sensors*, vol. 15, no. 1, pp. 2059–2085, 2015.

[43] F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, and M. Griss, *Activity-Aware Mental Stress Detection Using Physiological Sensors*, 2012, pp. 211–230.

[44] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, 2010.

[45] L. Wang, “Recognition of human activities using continuous autoencoders with wearable sensors,” *Sensors*, vol. 16, no. 2, 2016.

[46] W. Xu, X. Liu, and Y. Gong, “Document clustering based on non-negative matrix factorization,” in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, 2003, pp. 267–273.

[47] L. Zhang, X. Wu, and D. Luo, “Recognizing human activities from raw accelerometer data using deep neural networks,” in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 2015, pp. 865–870.

[48] ——, “Recognizing human activities from raw accelerometer data using deep neural networks,” in *14th IEEE International Conference on Machine Learning and Applications, ICMLA 2015, Miami, FL, USA, December 9-11, 2015*, 2015, pp. 865–870.

[49] M. Zhang and A. A. Sawchuk, “Human daily activity recognition with sparse representation using wearable sensors,” *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 3, pp. 553–560, 2013.

[50] ——, “Usc-had: A daily activity dataset for ubiquitous activity recognition using wearable sensors,” in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ser. UbiComp ’12, 2012, pp. 1036–1043.