Information Fusion 40 (2018) 45–56

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Multi-view stacking for activity recognition with sound and accelerometer data

Enrique Garcia-Ceja*, Carlos E. Galván-Tejada, Ramon Brena

Tecnologico de Monterrey, Av. Eugenio Garza Sada 2501 Sur, Monterrey, NL 64849, Mexico

ARTICLE INFO

Article history:
Received 5 December 2016
Revised 5 April 2017
Accepted 11 June 2017
Available online 12 June 2017

Keywords:
Multi-view learning
Stacked generalization
Activity recognition
Accelerometer
Sound

ABSTRACT

Many Ambient Intelligence (AmI) systems rely on automatic human activity recognition for getting crucial context information, so that they can provide personalized services based on the current users' state. Activity recognition provides core functionality to many types of systems including: Ambient Assisted Living, fitness trackers, behavior monitoring, security, and so on. The advent of wearable devices along with their diverse set of embedded sensors opens new opportunities for ubiquitous context sensing. Recently, wearable devices such as smartphones and smart-watches have been used for activity recognition and monitoring. Most of the previous works use inertial sensors (accelerometers, gyroscopes) for activity recognition and combine them using an aggregation approach, i.e., extract features from each sensor and aggregate them to build the final classification model. This is not optimal since each sensor data source has its own statistical properties. In this work, we propose the use of a multi-view stacking method to fuse the data from heterogeneous types of sensors for activity recognition. Specifically, we used sound and accelerometer data collected with a smartphone and a wrist-band while performing home task activities. The proposed method is based on multi-view learning and stacked generalization, and consists of training a model for each of the sensor views and combining them with stacking. Our experimental results showed that the multi-view stacking method outperformed the aggregation approach in terms of accuracy, recall and specificity.

1. Introduction

Ambient Intelligence (AmI) [1] is an emerging discipline that brings intelligence to our everyday environments by adapting those environments to our needs [1], making them aware of the context [2]. It builds upon advances in sensors, pervasive computing, and artificial intelligence. AmI technologies should be sensitive, responsive, adaptive, transparent, ubiquitous, and intelligent. In an AmI environment, devices are expected to work collectively by sharing information and using the history of past events. The AmI vision puts lighting, sound, vision, domestic appliances, and personal health care products to cooperate seamlessly in order to help the user [3].

Several of these systems are based on Human Activity Recognition (HAR) since knowing the current activity is of great importance to understand the users' context. Following Dey's notion of context [4], the user's task is one of its key elements. This is why HAR research has received great interest recently [5–13]. Being able to detect the activities being performed by an individual can provide valuable information in the process of understanding the context and situation in a given environment, so it is of great interest because of the wide range of possible applications such as in medicine, Ambient-Assisted Living [14], sports, marketing [15], surveillance [16], etc.

Data from several sources is collected and then analyzed to extract useful context information in many HAR applications, and some works have explored the combination of different types of sensors for activity recognition [5–7]. However, most of them use an *aggregation* approach, i.e., extract features from each sensor and aggregate them to train a predictive model. Aggregation is not optimal since each sensors' data have their own statistical properties [17] and combining them in the same model can confound them. In this paper, we present a method based on multi-view learning and stacked generalization for fusing audio and accelerometer sensor data for human activity recognition using wearable devices. We treat each sensor's data as different *views* and then, they are combined using stacked generalization [18]. The proposed approach is flexible since it does not rely on a specific classification model and is efficient in terms of memory and computation since it only requires to train a classifier for each sensor type and an ex-

* Corresponding author.
E-mail addresses: e.g.mx@ieee.org (E. Garcia-Ceja), ericgalvan@uaz.edu.mx (C.E. Galván-Tejada), ramon.brena@itesm.mx (R. Brena).

http://dx.doi.org/10.1016/j.inffus.2017.06.004
1566-2535/© 2017 Elsevier B.V. All rights reserved.

tra meta-classifier. We evaluated the multi-view stacking approach for home tasks activity recognition using sound and accelerometer data collected with a smartphone and a wrist-band. Furthermore, we evaluated the proposed method with other three multi modal HAR datasets.

This document is organized as follows: Section 2 presents an overview of activity recognition and sensor fusion methods that have been used. This Section also presents the background on multi-view learning and stacked generalization. In Section 3 we explain how stacked generalization is used in the context of multi-view learning for activity recognition. Section 4 details the accelerometer/audio data collection process. In Section 5 we explain the feature extraction process for the audio and accelerometer data. Next, in Section 6 the experiments and results are presented and finally in Section 7 we draw the conclusions.

2. Related work and background

There are two main types of sensors that have been used for Human Activity Recognition: *external sensors* and *wearable sensors*. External sensors are installed in the environment and may not have direct physical contact with the user. Examples of such sensors are: video cameras, microphones, motion sensors, depth cameras like the Microsoft Kinect, RFID tags, switches, etc. On the other hand, wearable sensors are carried by the user or are embedded in devices such as smartphones, smartwatches and fitness bracelets. Examples of wearable sensors are: accelerometers, gyroscopes, magnetometers, to name a few.

Automatic activity recognition systems have been successfully developed using external sensors such as video cameras [19–21] and color-depth cameras [8,9]. With the recent advent of smartphones and wearable devices such as smart-watches and fitness bands, it is now possible to collect data from their different sensors without the need of a fixed infrastructure. Recently, the sensors embedded in those type of devices have been used for human activity recognition given their flexibility, ubiquity and unobtrusiveness. Often, inertial sensors (accelerometers and gyroscopes) are used for HAR tasks, albeit other types of sensors like microphones, light sensors, temperature, heart rate, etc. are also embedded in those type of wearable devices. The use of wearable sensors [22] has gained a lot of attention because they have several advantages; in particular, the recognition can be performed in any place unlike video cameras in which it is restricted to a specific area. Another problem of external sensors is that in environments with multiple residents it becomes difficult to detect which person activated a specific sensor. This is not a problem for wearable sensors since they are personal. Yet other problems of external sensors are related to privacy issues, because the user does not decide if s/he is going to be monitored.

A common recent trend is to use smartphones for HAR since they are immensely popular and they already have several types of embedded sensors. Another advantage is that all the processing can be performed inside the phone so there is no need to carry another processing unit. One of the first works to perform all the recognition inside a phone was the one of Brezmes et al. [23]. There are also other works that have used smartphones for activity recognition [10–13]. Given the advantages of wearable sensors, in this work we focus on this type of systems. Specifically, we used a smartphone and a wrist-band to perform the recognition.

2.1. Sensor fusion in activity recognition

With the increasing miniaturization of sensors, it is now common to find many types of them in our environment, especially in wearable devices. Given their ubiquity and sensing capabilities, a wide range of physical phenomena can be measured, thus, generating large quantities of diverse data types. These data can be used to extract contextual information from the environment allowing the realization of reactive systems based on the current inferred state. Combining the diverse sources of data in an intelligent manner in order to generate knowledge and extract useful information has been an active research area [24]. In multimedia analysis, it is common to have different sensing modalities such as video, audio, text, WWW resources, etc. and the fusion of the multiple sources can increase the accuracy of the system [25]. In medical image analysis, the fusion of different imaging modalities (MRI, Ultrasound, CT, PET, SPECT) can produce improved results [26]. For human activity recognition, the most common approach is to use inertial sensors (accelerometers, gyroscopes, tilt switches, etc.) due to their flexibility and infrastructureless capabilities; however, external sensors (cameras, switches, motion sensors, RFID) are also used, mainly in smart environments.

In the work of Tolstikov et al. [27] Dynamic Bayesian Networks and Dempster-Shafer theory are used to fuse sensor data in order to recognize seven activities of daily living using 14 binary sensors placed around a house. Their results suggest that both methods are similar in terms of performance. Amoretti et al. [28] also used Bayesian Networks to monitor activities for ambient assisted living in a smart environment. In wearable sensor settings, the most common approach of sensor fusion is *aggregation*, i.e., concatenate the extracted features from all sensors and train a single classification model with them. Shoaib et al. [5] explored the use of smartphones' accelerometers and gyroscopes tested individually and in combination for activity recognition. They used an aggregation approach and found that the combination of both sensors improved the overall performance when the individual performances are not very high. In a later work [6] of the same authors (Shoaib et al.) they combined smartphone and wrist-worn inertial sensors by aggregation and–again– obtaining better results when fusing both devices. Dernbach et al. [7] also conducted experiments with accelerometer and gyroscope sensors and also concluded that combining information from both sensors (by aggregation) increased the system performance by 10–12% accuracy. In a similar work, Hayashi et al. [29] also used an aggregation of accelerometer and sound data to classify daily activities achieving better results when using both sources of information.

Even though it has been shown that combining multiple sources of data can increase the system accuracy, *aggregation* is not optimal since each sensor has its own statistical properties [17] which may require a different treatment. Zhu & Sheng [30] used a multi-sensor fusion scheme for combining sensors attached to the foot and waist. They trained two different neural networks for each of the sensors to recognize coarse-grained activities. Then, the outputs of the neural networks are fused using manually defined rules. These rules dictate if the fine-grained activity classification should be performed by heuristic discrimination or a Hidden Markov Model. By training two different neural networks they are able to preserve the statistical properties of each sensor unit; however, the manual definition of rules becomes hard when increasing the number of activities, thus, limiting the scalability of the approach. Another sensor fusion method was proposed by Banos et al. [31] whose aim is to be robust against hardware failures. Their proposed classifier is trained in a hierarchical fashion by first generating $m \times n$ binary classifiers where m is the number of sensors and n the number of classes, i.e., n binary classifiers for each sensor. Then, these classifiers are weighted and aggregated with a particular function for each sensor which corresponds to the second level classifiers. Finally, the decisions of the second level classifiers are weighted and combined to produce the final prediction. This method proved to be very robust in the presence of sensor failures by combining information of the

Table 1
Related work.

Reference	Type of sensors	Data sources	**Fusion method**
Tolstikov et al. [27]	external	binary sensors	Dynamic Bayesian Networks and Dempster-Shafer theory.
Amoretti et al. [28]	external	time-of-flight cameras, intelligent carpets, accelerometers	Bayesian Networks
Shoaib et al. [5]	wearable	accelerometers, gyroscopes	Aggregation
Shoaib et al. [6]	wearable	wrist-worn and smartphone sensors	Aggregation
Dernbach et al. [7]	wearable	accelerometers, gyroscopes	Aggregation
Hayashi et al. [29]	wearable	accelerometers, sound	Aggregation
Zhu & Sheng [30]	wearable	inertial sensors in waist and foot	Decision rules
Banos et al. [31]	wearable	bi-axial accelerometers	Hierarchical weighted classifier
Nishida et al. [32]	wearable	accelerometers, sound	Gaussian Mixture Models Weighting
This work.	wearable	accelerometer, sound	Multi-View Stacking

available working sensors, however, it requires a large number of classifiers. Nishida et al. [32] conducted experiments to recognize human activities using smartphone accelerometer data and sound recorded with a camera. They trained two Gaussian Mixture Models and evaluated different weights for the accelerometer data in order to vary its importance. The limitation of this approach is that it requires to find an additional weighting parameter and the number of mixtures to use. Table 1 presents a summary of representative activity recognition works classified by type of sensors, data sources and sensor fusion method.

This work differs from the previous ones in the following aspects: Firstly, it combines the data from heterogeneous types of sensors to complement each other and thus, increase recognition accuracy. Secondly, it is efficient in terms on the number of models to be trained since it only requires a classifier for each sensor and a meta-classifier; in comparison to other approaches like in [31] which requires to train a model for each class and for each sensor. Thirdly, it is flexible in the choice of classifiers, i.e., it can potentially include combinations of different types of models (this will be left as future work). Fourthly, we used a combination of wearable devices (smartphone and wrist-band) which are commonly used by many people in their everyday life, thus, reducing obtrusiveness issues. Finally, it is based on extensively studied machine learning methods namely: *multi-view learning* and *stacked generalization* which are detailed in the following sections.

2.2. Multi-view learning

It is not unusual to have applications in which each observation can be represented by different sets of features or 'views'. For example, a video can be represented by the information contained in its sequence of images but also in the audio itself. A web-page can be characterized by the text contained within it but also by the hyperlinks pointing to that page. For machine learning tasks, features from the different views can be simply aggregated to learn a given model. This approach might not be optimal since each view has its own statistical properties [17]. Another paradigm called Multi-view learning deals with the problem of learning a model based on the different views of the data [33,34]. One of the earliest works in this direction is the one of Blum & Mitchell [35] which was developed for semi-supervised learning tasks [36], i.e., when there are large amounts of unlabeled instances. They considered the problem of web-page classification with two views: the text in the web-page and the hyperlinks pointing to it. Their co-training method consists of initially training two independent classifiers (one for each view) and then perform several iterations. In each iteration, one of the classifiers labels a subset of the unlabeled instances and the instances with the most confident predictions are added as training data to the other classifier and vice versa. In this way, the classifiers help each other by augmenting their training set to make use of the unlabeled data. This approach assumes that each view is sufficient to train a good classifier and that both views are conditionally independent given the class. Zhou & Li relaxed those assumptions by introducing a tri-training method which uses three classifiers [37]. Sometimes, the data cannot be naturally represented by different views and thus, some approaches aim to synthetically construct the different representations [38,39]. The previous mentioned works were developed in the context of semi-supervised learning. To a lesser extent, there have also been works in multi-view learning for supervised learning. For example, Farquhar et al. [40] proposed a method that combines kernel Canonical Correlation Analysis and a Support Vector Machine for image classification obtaining accuracy improvements compared to using individual SVMs. Diethe et al. [41] extended Fisher discriminant analysis classification to the multi-view case and later, Chen & Sun proposed a hierarchical multi-view Fisher discriminant analysis method [42]. In a recent work, Wang et al. [43] proposed a linear multi-view classifier based on *intact* feature vectors. The approach assumes that the different views of an observation are generated from one single intact vector which is recovered guiding the search using label information.

In this work we will consider two different representations of the activities from different sources, namely: accelerometer and audio sensors. We will use a multi-view learning approach by considering each source of information independent of the other and fusing them using *Stacked Generalization* which is described in the next section and the proposed multi-view stacking method is detailed in Section 3.

2.3. Stacked generalization

The concept of *stacked generalization* (also called *stacking*) was introduced by Wolpert (1992) [18] and is a type of ensemble method for combining multiple learners. The method consists of training a set of learners (called the *first-level learners*) with the original training data. The outputs of the first-level learners are then used to train a second-level learner called the *meta-learner*. For a basic introduction to ensemble learning and stacking see a textbook like Kubat's one [44] and Zhou [45]. The overall procedure comprises the following steps:

1. Define a set $\mathscr{L}$ of first-level learners and a *meta-learner*.
2. Train the first-level learners in $\mathscr{L}$ using the original training data $\mathbf{D}$ which contains n instances.
3. Predict the labels of $\mathbf{D}$ with each of the learners in $\mathscr{L}$; each of the $|\mathscr{L}|$ learners gives a prediction vector $\mathbf{p}$ of n elements.
4. Form a new matrix $\mathbf{M}_{n\times|\mathscr{L}|}$ by column binding the prediction vectors and the true labels $\mathbf{y}$ to produce the new training data $\mathbf{D}'$
5. Train the *meta-learner* with $\mathbf{D}'$
6. Output the final stacking model $\mathcal{S} :< \mathscr{L}, \textit{meta-learner} >$.

Fig. 1 shows the procedure to generate the new training data $\mathbf{D}'$ for the meta-learner.

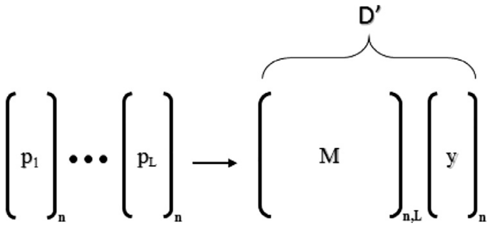


Fig. 1. Depiction of the process to produce the new training data D' for the meta-learner by column-binding the predictions of the learners in $\mathscr{L}$ and the true labels **y**.

In steps 2 and 3, there is a high risk of over-fitting since the predictions are made with the same data used to train the models. To avoid this, steps 2 and 3 are usually performed using k-fold cross validation. After $\mathbf{D}'$ has been generated, the learners in $\mathscr{L}$ are retrained using all instances in **D**.

In the original work of Wolpert [18], stacked generalization was used for classification and surface-fitting. Later, it was also used for regression by stacking regression trees [46] and for unsupervised learning to estimate densities [47]. In the context of classifications tasks, Ting & Witten [48] showed that adding confidence information about the predictions for the meta-learner can lead to better classification results. In this work (Section 3) we will use stacked generalization as a means to fuse the different activities' views in order to generate the final classifier.

3. Multi-view stacking

The proposed multi-view stacking classification approach consists of training one first-level learner for each view and combining their outputs using stacked generalization; in our case one view will comprise the information coming from the accelerometers, and the other one coming from the sound. The base classifiers for each of the views will take as data the set of features resulting from either accelerometers or sound (see Section 5). The dataset D' that will serve to train the meta-learner is generated by column binding the outputs of each of the first-level learners and the true labels **y**. These outputs consist of the predicted labels and the associated predicted probabilities for each of the k possible classes. The output probabilities of the first-level learners are averaged. Thus, the final feature vectors have size $|\mathscr{L}|+k+1$, with the form $[l_1, .., l_i, .., l_{|\mathscr{L}|}, p_1, .., p_i, .., p_k, y]$ where l_i is the predicted label of each first-level learner, p_i are the averaged probabilities for each possible class k and y is the true label.

In Stacked Generalization, algorithms that in general produce high performance results are used as first-level learners such as neural networks, Support Vector Machines (SVMs), Random Forests, etc. For our experiments, we used Random Forests [49] since they have been shown to produce good overall results [50] and in particular, they have also proven to outperform other classifiers in HAR tasks [51–54]. In the work of Casale et al. [51], Random Forest outperformed decision tree, Bagging and Boosting classifiers in recognizing five different activities individually and overall. Weiss and Lockhart [52] tested 8 different classification algorithms for HAR including: Naive Bayes, Neural Networks, instance based learning, among others, and on average, Random Forest produced the best results. Nguyen et al. [53] experimented with several classifiers for HAR, including k-NN and SVM and they reported that Random Forest consistently achieved the best results. In the work of Galván-Tejada et al. [54], they used sound data for HAR and obtained the best results when using Random Forest compared to Neural Networks.

As opposed to other multi-view learning algorithms that are tied to a specific learner, stacked generalization has the advantage that any type of models can be used as a first-level learners and meta-learners, providing more flexibility for implementation. Often, heterogeneous types of first-level learners are used, thus, providing more diversity and adaptation for each of the views, e.g., an optimization method can be used to select the subset of best learners for the given task [55]. In order to make the comparison between only audio, only accelerometer, aggregated data and multi-view stacking as fair as possible, we used for all cases random forest as the first-level learners and also as the meta-learner (see Section 6).

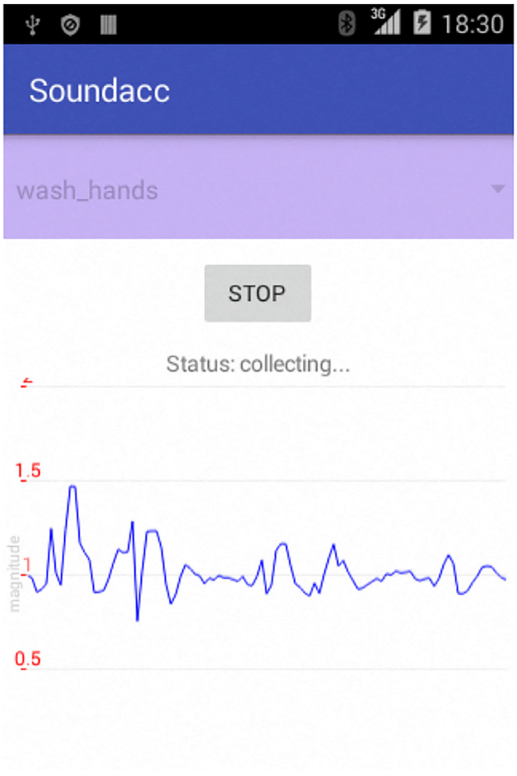


Fig. 2. Data collection cellphone application.

4. Data collection

For our home tasks activities dataset, the sound and accelerometer data were collected by 3 volunteers while performing 7 different activities: *mop floor, sweep floor, type on computer keyboard, brush teeth, wash hands, eat chips* and *watch t.v.*. Each volunteer performed each activity for approximately 3 min. If the activity lasted less than 3 min, another session was recorded until completing the 3 min. The data were collected with a wrist-band (Microsoft Band 2) and a cellphone. The wrist-band was used to collect accelerometer data and was worn by the volunteers in their dominant hand. The accelerometer sensor returns values from the *x, y* and *z* axes and the sampling rate was set to 31 Hz. The cellphone was used to record environmental sound with a sampling rate of 8000 Hz and it was placed on a table in the same room the user was performing the activity. An application for the Android operating system was developed to collect the data (Fig. 2). The application has a dropdown list from which the users can select the activity and a chart to display the accelerometer magnitude while recording. The wrist-band sends the sensor readings and a timestamp via Bluetooth to the cellphone and they are stored as text files.

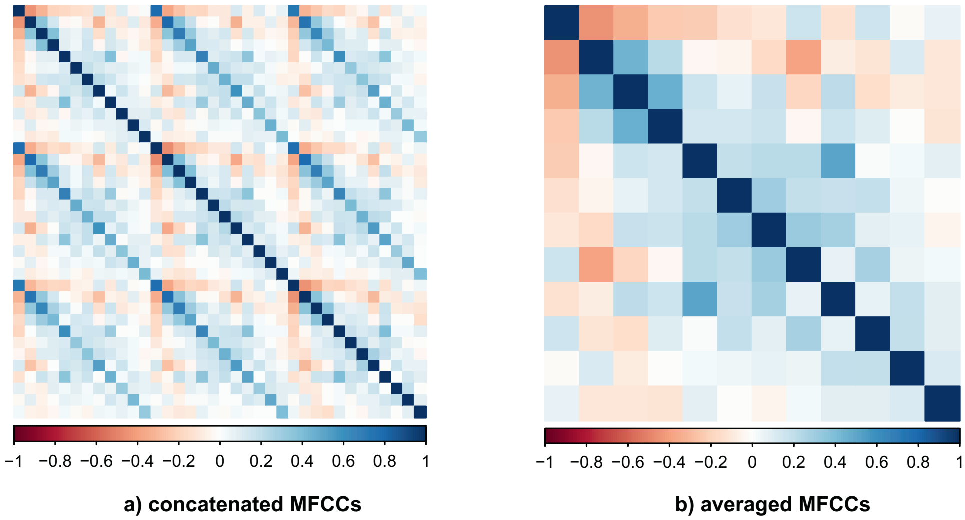


Fig. 3. Correlations between audio features: a) concatenated MFCCs b) averaged MFCCs.

5. Feature extraction

The feature extraction process consists of computing representative measures from the original signal in order to have a more compact representation while still preserving its discriminative characteristics. The original accelerometer and audio signals were segmented into fixed length windows of 3 s each, with no overlap since according to Banos et al. [56] this is the typical value for activity recognition systems, and in their extensive evaluation of different window lengths, they showed that small window sizes lead to better results than using longer window sizes. Characteristic measures (features) are then computed for each window segment. The resulting set of features for each segment is referred to as *feature vector* or an *instance*. Each *instance* will be represented by two sets of features corresponding to the different *views*: The acceleration view and the sound view. Next, we describe the extracted features from both, accelerometer and audio signals.

5.1. Accelerometer features

From the raw accelerometer signals, 16 features were extracted: The *mean* value of each of the 3 axes, the *standard deviation* of each of the 3 axes, the *max* value of each of the 3 axes, the *correlation* between each pair of axes, the *mean magnitude*, the *standard deviation of the magnitude*, the *magnitude area under the curve* (AUC, Eq. (1)) , and *magnitude mean differences* between consecutive readings (Eq. (2)). The *magnitude* of the signal represents the overall contribution of acceleration of the 3 axes (Eq. (3)). These type of features were chosen because they have shown to produce good results for activity recognition tasks [7,57,58].

$$AUC = \sum_{t=1}^{T} magnitude(t) \tag{1}$$

$$meandif = \frac{1}{T-1}\sum_{t=2}^{T} magnitude(t) - magnitude(t-1) \tag{2}$$

$$Magnitude(x, y, z, t) = \sqrt{a_x(t)^2 + a_y(t)^2 + a_z(t)^2}, \tag{3}$$

where $a_x(t)^2$, $a_y(t)^2$ and $a_z(t)^2$ are the squared accelerations at time t and T is the last time interval.

Table 2
Distribution of activities by class.

Class	Proportion
Brush teeth	12.98%
Eat chips	20.34%
Mop floor	13.05%
Sweep	12.84%
Type on keyboard	12.91%
Wash hands	12.98%
Watch t.v.	14.90%

5.2. Audio features

To characterize each audio signal, we extracted their Mel Frequency Cepstral Coefficients (MFCCs) since they have proven to produce good results for activity recognition [29,32,54,59].

Each 3 s audio segment was divided into three 1 s subsegments. From each 1 s sub-segment, 12 Mel Frequency Cepstral Coefficients (MFCCs) were computed, thus, resulting in a total of 36 MFCCs. The total number of instances were 1386 of 3 s each. The computation was performed using the R tuneR package [60]. One way to get the final feature vector of the entire 3 s segment is to concatenate the MFCCs but doing so may result in highly correlated features, i.e., coefficient 1 will be highly correlated with coefficient 13, 2 with 14 and so on. Fig. 3-a shows the correlations plot when concatenating the MFCCs. Here, we can see many correlation patterns (the blue diagonal lines). When building classification models, it is desirable to avoid highly correlated features. In order to avoid this correlations, we opted to average the MFCCs instead of concatenating them. Fig. 3-b shows the correlations plot when averaging the MFCCs. Here we can see that there are no visible strong correlation patterns as before. Another advantage of averaging is that the total number of features is reduced from 36 to just 12.

All the features were normalized between $[0-1]$, inclusive. Table 2 shows the distribution by class. Here, we can see that there is no considerable class imbalance.

To visualize how well the features of the two views (accelerometer and sound) can discriminate between classes, their first 3 coordinates after applying a multidimensional scaling (MDS) transformation [61] were plotted (Fig. 4). We also plotted the coordinates when aggregating the features of both views. The accelerometer view seems to have more compact and defined groups than the

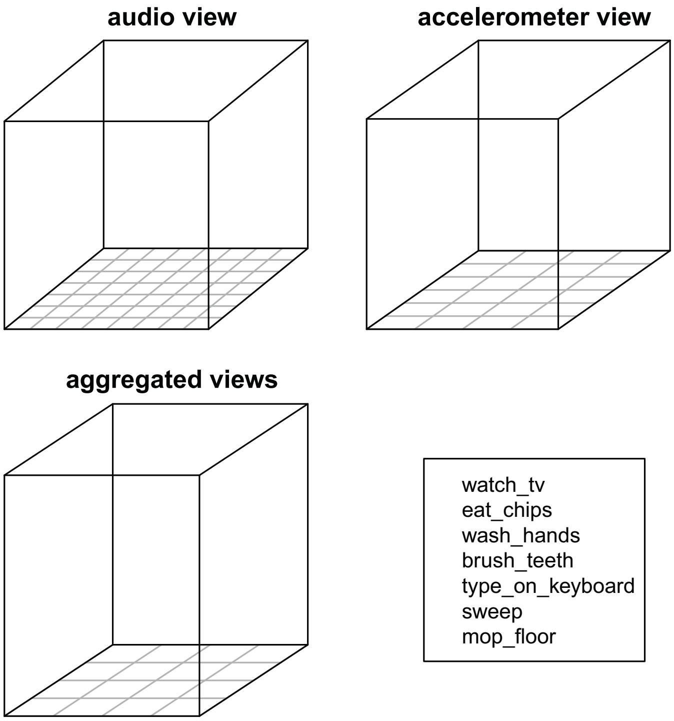


Fig. 4. Multidimensional scaling for the different views.

audio view. For example, the *watch tv* activity (yellow) seems to form a single group in the accelerometer view, whereas in the audio view it looks more fragmented. When aggregating both views, distinguishable groups can be identified, specially the *watch tv, wash hands* and *type on keyboard* activities. This exploratory analysis suggests that these features have the potential to capture the discriminative information of the different activities.

6. Experiments and results

To evaluate the proposed multi-view stacking approach for activity classification, four different configurations were considered for our home tasks activities dataset:

- **Audio view.** Perform the classification with just audio features.
- **Accelerometer view.** Perform the classification with just accelerometer features.
- **Aggregated views.** Perform the classification by concatenating both, audio and accelerometer features.
- **Multi-View Stacking.** Perform the classification using the proposed approach by building individual models for each view and combining them using stacked generalization.

Table 3
Performance metrics results for home tasks dataset. Average (standard deviation).

	Accuracy	Recall	Specificity
Audio view	0.838 (0.019)	0.836 (0.021)	0.972 (0.003)
Accelerometer view	0.854 (0.024)	0.844 (0.022)	0.975 (0.004)
Aggregated views	0.921 (0.026)	0.915 (0.031)	0.986 (0.004)
Multi-View Stacking	**0.941 (0.024)**	**0.939 (0.027)**	**0.990 (0.004)**

A random forest classifier was used for the four configurations. 10 fold cross validation was used to evaluate each configuration and the averaged performance metrics were reported. For multi-view stacking, 10 fold cross validation was used on the training data to build the dataset D' for the meta-learner. The reason to perform k-fold cross validation to generate the predictions is to avoid overfitting the training data $\mathbf{D}'$. At the end, in order to build the final stacked model $\mathcal{S}$, the $\mathscr{L}$ learners are retrained with all the training data $\mathbf{D}$.

The following performance metrics for each configuration were computed:

- *Accuracy:* This refers to the proportion of correctly classified instances.
- *Sensitivity or recall (true positive rate):* The proportion of positives that are correctly classified as such.
- *Specificity (true negative rate):* The proportion of negatives that are correctly classified as such.

Table 3 shows the results for each configuration and metric. Here, we can see that the accelerometer performed better than the audio and the performance was boosted when aggregating both views, specially in terms of accuracy and recall. The best performance was achieved with the proposed multi-view stacking method for the three metrics. Fig. 5 shows the resulting boxplots for the accuracy, recall and specificity between multi-view stacking and aggregated views. Clearly, the performance of multi-view stacking outperformed the approach of just aggregating all the views' features. A Wilcoxon signed rank test was used to test for statistically significant increase with $\alpha = 0.05$. Table 4 shows the tests' results. In all cases the difference was statistically significant.

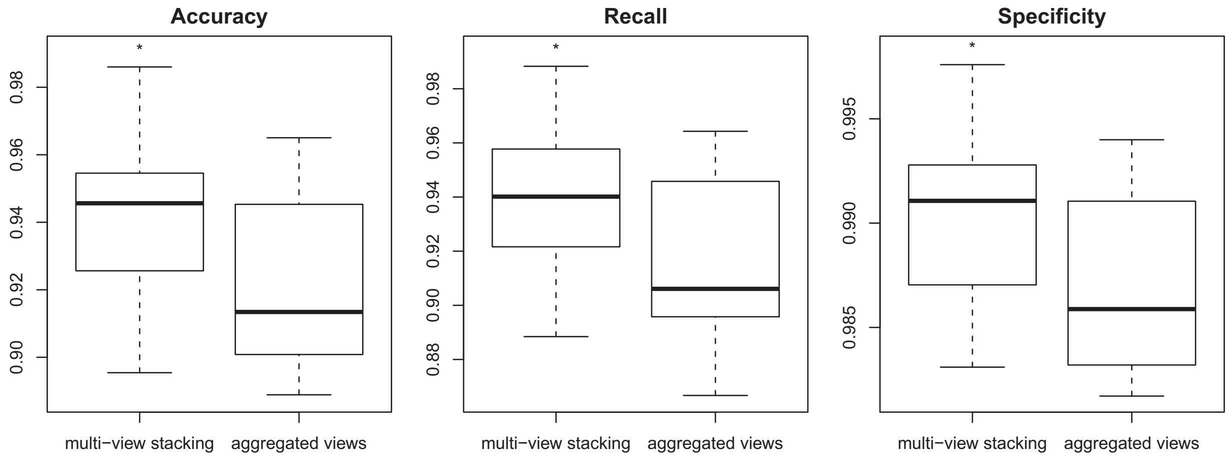
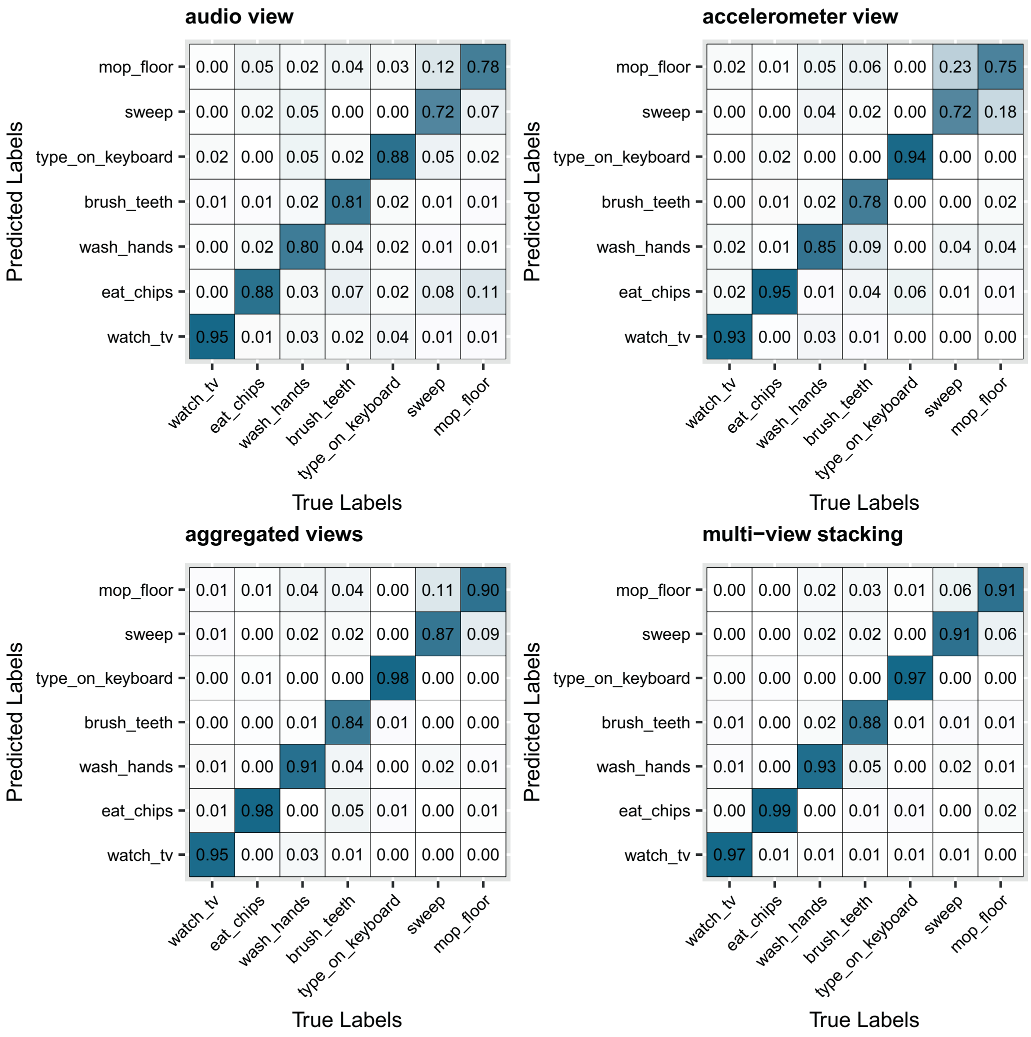


Fig. 5. Home tasks dataset boxplots (* statistically significant).



audio view

	watch_tv	eat_chips	wash_hands	brush_teeth	type_on_keyboard	sweep	mop_floor
mop_floor	0.00	0.05	0.02	0.04	0.03	0.12	0.78
sweep	0.00	0.02	0.05	0.00	0.00	0.72	0.07
type_on_keyboard	0.02	0.00	0.05	0.02	0.88	0.05	0.02
brush_teeth	0.01	0.01	0.02	0.81	0.02	0.01	0.01
wash_hands	0.00	0.02	0.80	0.04	0.02	0.01	0.01
eat_chips	0.00	0.88	0.03	0.07	0.02	0.08	0.11
watch_tv	0.95	0.01	0.03	0.02	0.04	0.01	0.01

accelerometer view

	watch_tv	eat_chips	wash_hands	brush_teeth	type_on_keyboard	sweep	mop_floor
mop_floor	0.02	0.01	0.05	0.06	0.00	0.23	0.75
sweep	0.00	0.00	0.04	0.02	0.00	0.72	0.18
type_on_keyboard	0.00	0.02	0.00	0.00	0.94	0.00	0.00
brush_teeth	0.00	0.01	0.02	0.78	0.00	0.00	0.02
wash_hands	0.02	0.01	0.85	0.09	0.00	0.04	0.04
eat_chips	0.02	0.95	0.01	0.04	0.06	0.01	0.01
watch_tv	0.93	0.00	0.03	0.01	0.00	0.00	0.00

aggregated views

	watch_tv	eat_chips	wash_hands	brush_teeth	type_on_keyboard	sweep	mop_floor
mop_floor	0.01	0.01	0.04	0.04	0.00	0.11	0.90
sweep	0.01	0.00	0.02	0.02	0.00	0.87	0.09
type_on_keyboard	0.00	0.01	0.00	0.00	0.98	0.00	0.00
brush_teeth	0.00	0.00	0.01	0.84	0.01	0.00	0.00
wash_hands	0.01	0.00	0.91	0.04	0.00	0.02	0.01
eat_chips	0.01	0.98	0.00	0.05	0.01	0.00	0.01
watch_tv	0.95	0.00	0.03	0.01	0.00	0.00	0.00

multi−view stacking

	watch_tv	eat_chips	wash_hands	brush_teeth	type_on_keyboard	sweep	mop_floor
mop_floor	0.00	0.00	0.02	0.03	0.01	0.06	0.91
sweep	0.00	0.00	0.02	0.02	0.00	0.91	0.06
type_on_keyboard	0.00	0.00	0.00	0.00	0.97	0.00	0.00
brush_teeth	0.01	0.00	0.02	0.88	0.01	0.01	0.01
wash_hands	0.01	0.00	0.93	0.05	0.00	0.02	0.01
eat_chips	0.00	0.99	0.00	0.01	0.01	0.00	0.02
watch_tv	0.97	0.01	0.01	0.01	0.01	0.01	0.00

Fig. 6. Confusion Matrices for home tasks dataset.

Table 4
Wilcoxon signed rank test results between multi-view stacking and aggregated views.

Performance metric	p-value
Accuracy	0.01
Recall	0.005
Specificity	0.005

Fig. 6 shows the resulting confusion matrices (in percentages). The antidiagonal of the matrices represents the recall of each individual activity. The multi-view stacking had a recall increase for all activities with respect to aggregated views except for the *type on keyboard* activity. The audio features were better at detecting the *watch t.v., brush teeth* and *mop floor* activities whereas the accelerometer features were better at detecting *eat chips, wash hands* and *type on keyboard*. For all configurations, the greatest error was between the *sweep* and *mop floor* activities.

6.1. Other datasets

To test the applicability of the proposed approach in different scenarios, we conducted experiments with other HAR datasets with similar characteristics (with an accelerometer in the wrist complemented with other sensors). For the purpose of comparison, the same set of features were extracted for all inertial sensors (see Section 5.1) and the same set of features were extracted for all audio sources (see Section 5.2). For datasets containing 3D skeleton data representations, the features were extracted by computing the distance between a reference joint point (the spine) and every other joint point for each frame [62] and taking the *mean, max* and *min* values across all time frames.

6.1.1. Berkeley MHAD dataset

The Berkeley MHAD dataset [63] consists of temporally synchronized and geometrically calibrated data from microphones, accelerometers, an optical motion capture system, multiple stereo cameras and depth sensors. The aim of this database is to provide researchers a benchmark to test new algorithms across multiple modalities. The data was captured by 12 subjects and contains 11 actions. All participants performed 5 repetitions for each action which are: 1-jumping in place, 2-jumping jacks, 3-bending, 4-punching, 5-waving two hands, 6-waving one hand, 7-clapping, 8-throwing a ball, 9-sit/stand up, 10-sit down and 11-stand up. The total number of recordings were 660. Due to some missing sensor data, 2 recordings were lost yielding a total of 658 instances. For our experiments we considered 3 different views: wrist-acceleration, audio and 3D skeleton points which are obtained from the video motion capture systems. Table 5 shows the obtained results for each of the 3 views independently, aggregated and with multi-view stacking. This table also presents the results

Table 5
Performance metrics results for Berkeley MHAD dataset. Average (standard deviation).

	Accuracy	Recall	Specificity
Audio view	0.682 (0.063)	0.698 (0.074)	0.968 (0.006)
Accelerometer view	0.954 (0.031)	0.957 (0.031)	0.995 (0.003)
Skeleton view	0.960 (0.027)	0.963 (0.024)	0.996 (0.002)
Aggregated views	0.987 (0.013)	0.987 (0.016)	0.998 (0.001)
Multi-View Stacking	**0.995 (0.007)**	**0.995 (0.007)**	**0.999 (0.0007)**
Reported in original work of Ofli et al. [63]	Accuracies from 0.938 (motion capture + depth data), 0.974 (motion capture + acc + audio), 1.0 (all sensors).		

Table 6
Performance metrics results for UTD-MHAD dataset. Average (standard deviation).

	Accuracy	Recall	Specificity
Accelerometer view	0.902 (0.034)	0.907 (0.034)	0.996 (0.001)
Gyroscope view	0.852 (0.039)	0.857 (0.050)	0.994 (0.001)
Skeleton view	0.909 (0.046)	0.917 (0.044)	0.996 (0.001)
Aggregated views	0.975 (0.017)	0.975 (0.022)	0.999 (0.0006)
Multi-View Stacking	**0.981 (0.016)**	**0.984 (0.015)**	**0.999 (0.0006)**
Reported in original work of Chen et al. [64], [65]	Accuracy 0.791 [64], and 0.972 [65]		

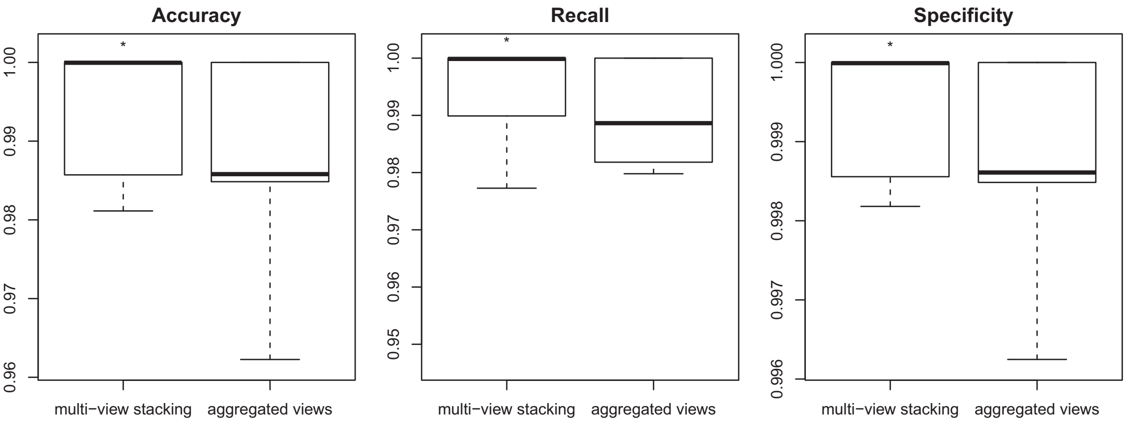


Fig. 7. Berkeley MHAD dataset boxplots (* statistically significant).

obtained in the original work. Here, we can see that the best performance was achieved with multi-view stacking. In the original work, Ofli et al. [63] reported an accuracy of 0.974 when using the same set of sensors (motion capture + accelerometer + audio) compared with the 0.995 accuracy that we obtained. In the original work, they achieved a 1.0 accuracy when combining all sensors which is very close to the 0.995 accuracy we achieved but just using data from motion capture, audio and the right wrist accelerometer.

Fig. 7 shows the resulting boxplots for the performance metrics. For the three cases, the increased performance of multi-view stacking compared with aggregating views was statistically significant. Fig. 8 shows the resulting confusion matrices for each view, aggregated views and multi-view stacking. From these matrices, we can see that the recall (anti-diagonal) of all activities was >= for multi-view stacking compared with aggregated views. Next, we present our results with another multi modal dataset.

6.1.2. UTD-MHAD dataset

The UTD-MHAD database [64] was collected using a Microsoft Kinect camera and a wearable inertial sensor with 3-axis accelerometer and a 3-axis gyroscope. The database has 27 actions performed by 8 subjects with 4 repetitions per action. Since there were 3 corrupted sequences, the total number of instances were 861. The actions include: 1-swipe left, 2-swipe right, 3-wave, 4-clap, 5-throw, 6-arm cross, 7-basketball shoot, 8-draw x, 9-draw circle CW, 10-draw circle CCW, 11-draw triangle, 12-bowling, 13-boxing, 14-baseball swing, 15-tennis swing, 16-arm curl, 17-tennis serve, 18-push, 19-knock, 20-catch, 21-pickup throw, 22-jog, 23-walk, 24-sit 2 stand, 25-stand 2 sit, 26-lunge and 27-squat. For activities 1–21, the inertial sensor was placed on the right wrist and for activities 22–27, it was placed on the right thigh. For our experiments, we considered three different views: the accelerometer, gyroscope and the skeleton produced by the Kinect sensor. Table 6 shows the obtained results. Again, the best accuracies were obtained with the multi-view stacking method. In the original work, the authors achieved an accuracy of 0.791 [64] and in a follow up work [65] they achieved an accuracy of 0.972 whereas the multi-view stacking had an accuracy of 0.981. Fig. 9 shows the resulting boxplots in which we can see that the performance of multi-view stacking was higher than that of aggregated views, though, not statistically significant. Fig. 10 shows the corresponding confusion matrices.

6.1.3. Opportunity dataset

This dataset consists of daily activities recorded with multi modal sensors [66]. The available database [67] contains recordings captured by 4 subjects. We considered the four locomotion

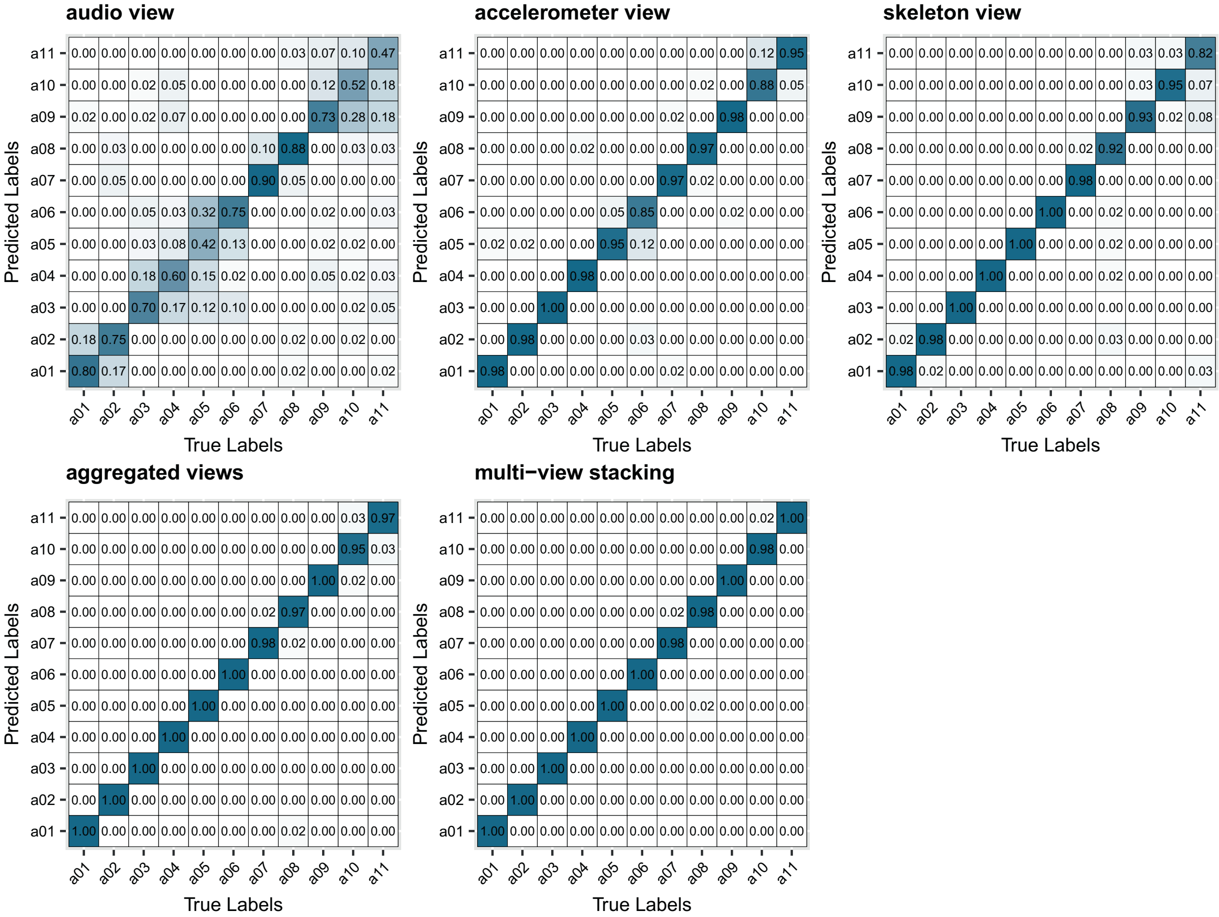


Fig. 8. Confusion Matrices for Berkeley MHAD dataset.

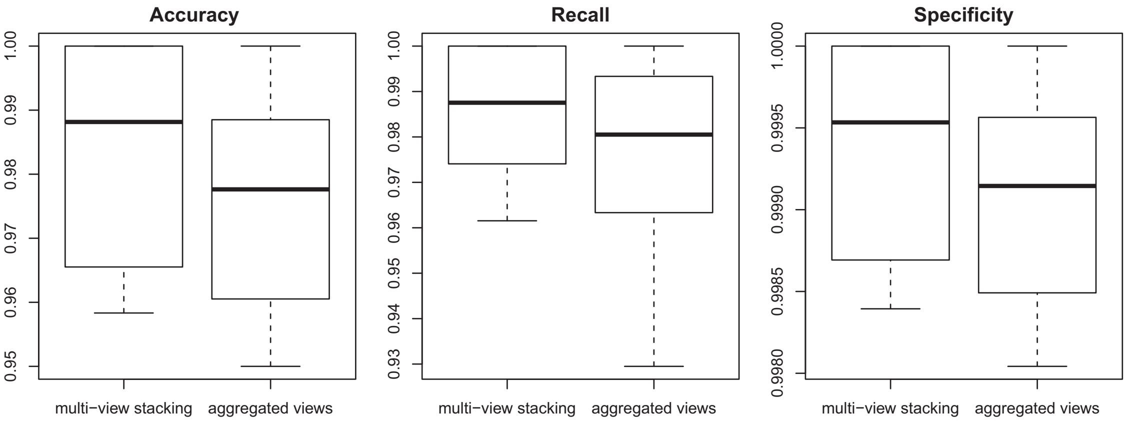


Fig. 9. UTD-MHAD dataset boxplots (* statistically significant).

activities included in the database: 1-stand, 2-walk, 3-sit and 4-lie. The total number of instances is 2477. We used the right wrist accelerometer, gyroscope and magnetometer as the three different views. Table 7 shows the results. In this case, multi-view stacking had the highest performance and also outperformed the accuracy reported in Sagha et al. [68]. Figs. 11 and 12 depict the resulting boxplots and confusion matrices.

From the experiments performed in our home tasks dataset and the other 3 benchmark datasets, we can see a similar behaviour:

Table 7
Performance metrics results for Opportunity dataset. Average (standard deviation).

	Accuracy	Recall	Specificity
Accelerometer view	0.843 (0.037)	0.790 (0.066)	0.925 (0.016)
Gyroscope view	0.821 (0.025)	0.692 (0.043)	0.914 (0.011)
Magnetometer view	0.889 (0.024)	0.855 (0.051)	0.948 (0.012)
Aggregated views	0.914 (0.020)	0.862 (0.036)	0.957 (0.009)
Multi-View Stacking	**0.925 (0.026)**	**0.905 (0.043)**	**0.965 (0.011)**
Reported in Sagha et al. [68]	Average accuracy of 0.83		

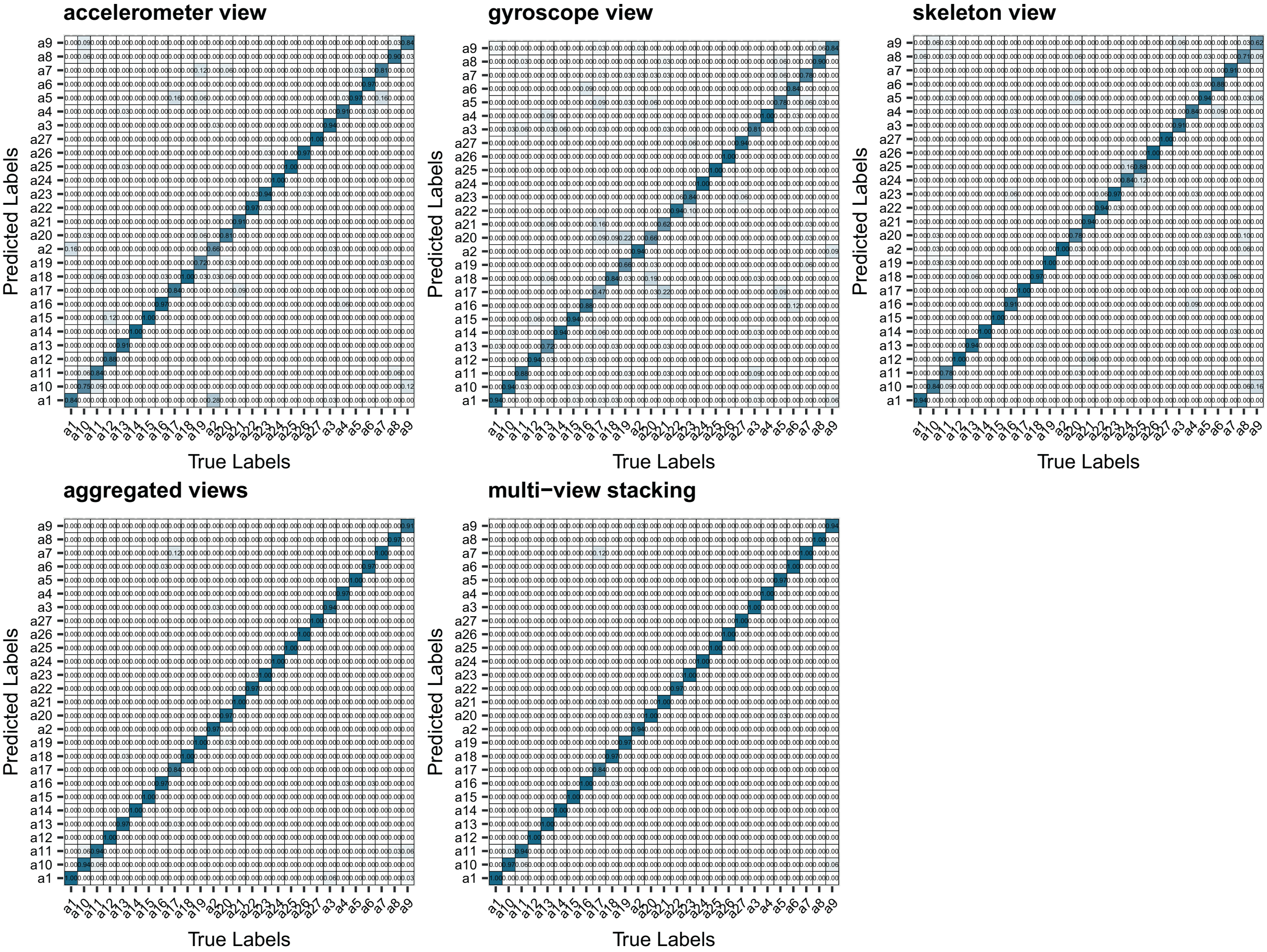


Fig. 10. Confusion Matrices for UTD-MHAD dataset.

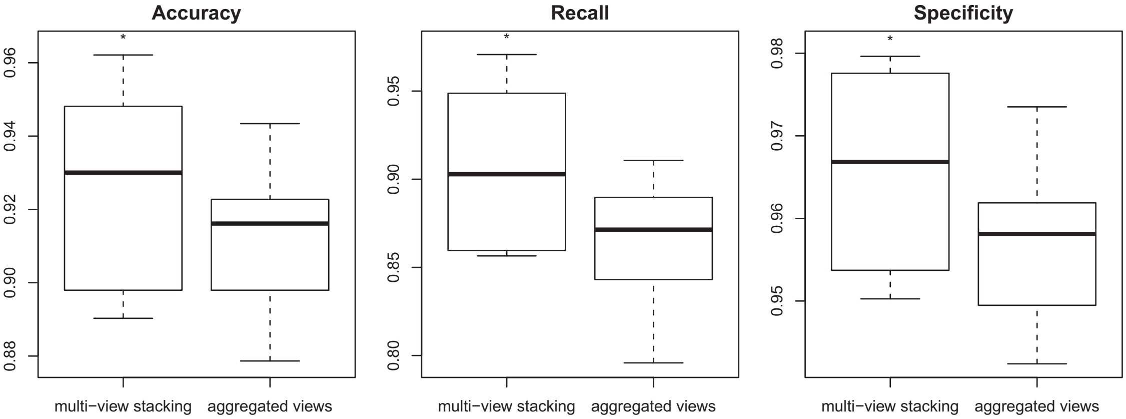


Fig. 11. Opportunity dataset boxplots (* statistically significant).

The performance metrics are higher when combining the different sensor views compared to using each sensor view independently. Furthermore, multi-view stacking produced better results than the aggregated views approach. A similar trend on the confusion matrices can also be observed across the different datasets. The anti-diagonal of the multi-view stacking confusion matrix has higher recall values than the other confusion matrices.

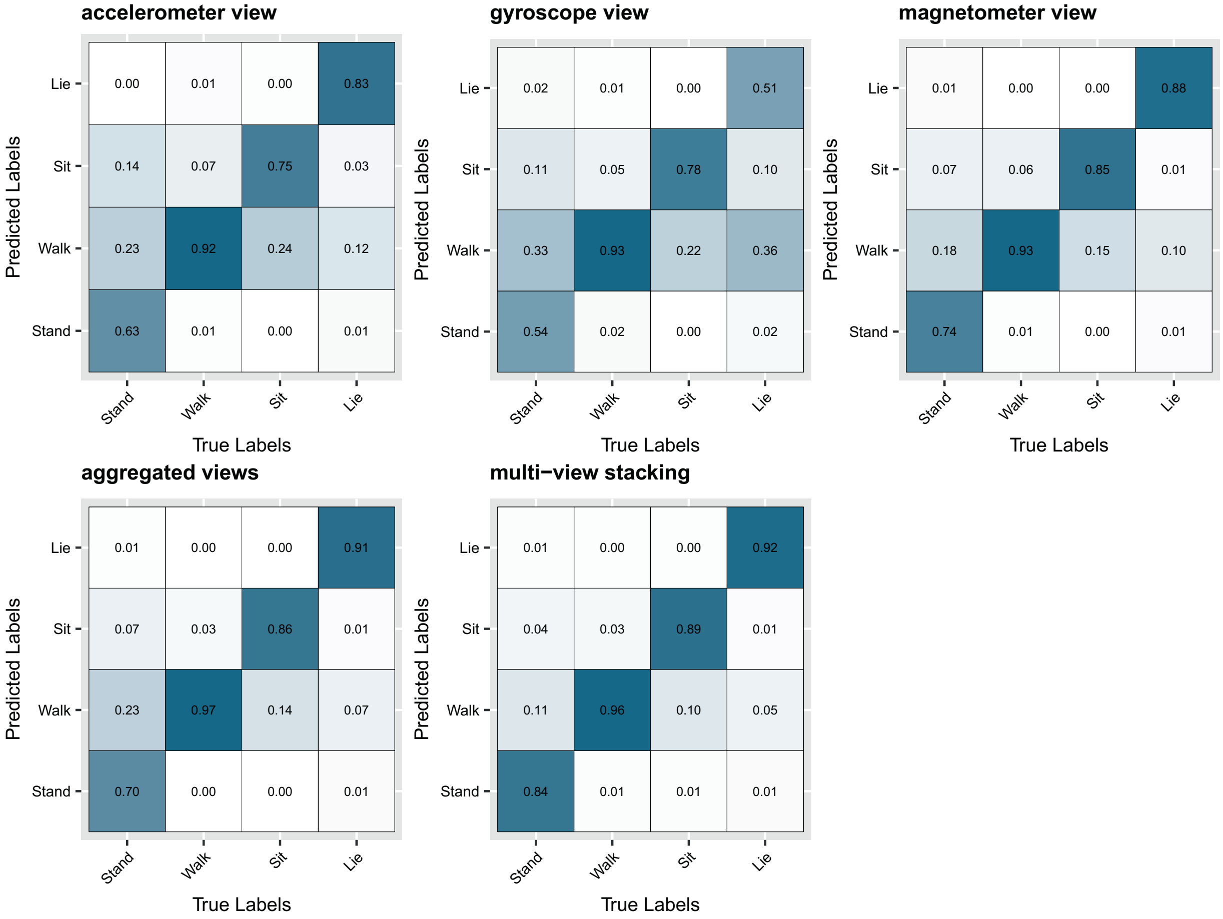


Fig. 12. Confusion Matrices for Opportunity dataset.

7. Conclusions

In this work we presented a method to fuse different types of sensors for activity recognition using wearable devices. We used sound and accelerometer data collected with a smartphone and a wrist-band for common home task activities. The proposed method is based on multi-view learning and stacked generalization. Each sensor was modeled as an independent view and the views were combined by stacking. Our results showed that the multi-view stacking method achieved better results than feature aggregation in terms of accuracy, recall and specificity. The experimental results also showed that combining sound and accelerometer data boosted the classification performance compared to using just one source of information. To validate the applicability of the proposed approach, we performed experiments with other 3 multi modal sensor HAR datasets obtaining similar results. Although these results are preliminary, they showed the potential of combining different types of sensors for activity recognition, particularly using multi-view and stacking methods. There are still several interesting problems to be explored; one of them is how to deal with missing data. This situation can arise due to sensor failure or because the user may decide to disable some sensors due to privacy concerns or to reduce battery consumption. A recognition system should be able to dynamically adapt itself to such scenarios. Another interesting future direction is to explore methods for finding the optimal combination of types of classifiers. Each sensors' data may be better modeled by specific base classifiers. The optimal sensor-classifier mapping could be found by using optimization methods such as Genetic Algorithms.

References

[1] D.J. Cook, J.C. Augusto, V.R. Jakkula, Ambient intelligence: technologies, applications, and opportunities, Pervasive Mob. Comput. 5 (4) (2009) 277–298.

[2] G.D. Abowd, A.K. Dey, P.J. Brown, N. Davies, M. Smith, P. Steggles, Towards a better understanding of context and context-awareness, in: International Symposium on Handheld and Ubiquitous Computing, Springer, 1999, pp. 304–307.

[3] E. Aarts, R. Wichert, Ambient intelligence, in: H.-J. Bullinger (Ed.), Technology Guide, Springer Berlin Heidelberg, 2009, pp. 244–249.

[4] A.K. Dey, Understanding and using context, Pers. Ubiquitous Comput. 5 (1) (2001) 4–7.

[5] M. Shoaib, S. Bosch, O.D. Incel, H. Scholten, P.J.M. Havinga, Fusion of smartphone motion sensors for physical activity recognition, Sensors 14 (6) (2014) 10146–10176, doi:10.3390/s140610146.

[6] M. Shoaib, S. Bosch, O.D. Incel, H. Scholten, P.J.M. Havinga, Complex human activity recognition using smartphone and wrist-worn motion sensors, Sensors 16 (4) (2016) 426, doi:10.3390/s16040426.

[7] S. Dernbach, B. Das, N.C. Krishnan, B.L. Thomas, D.J. Cook, Simple and Complex Activity Recognition through Smart Phones, in: Intelligent Environments (IE), 2012 8th International Conference on, 2012, pp. 214–221, doi:10.1109/IE.2012.39.

[8] J. Sung, C. Ponce, B. Selman, A. Saxena, Human activity detection from RGBD images, CoRR abs/1107.0169 (2011).

[9] R.F. Brena, A. Nava, Activity recognition in meetings with one and two kinect sensors, in: Mexican Conference on Pattern Recognition, Springer International Publishing, 2016, pp. 219–228.

[10] J.R. Kwapisz, G.M. Weiss, S.A. Moore, Activity recognition using cell phone accelerometers, SIGKDD Explor. Newsl. 12 (2) (2011) 74–82, doi:10.1145/1964897.1964918.

[11] Y.-S. Lee, S.-B. Cho, Activity Recognition Using Hierarchical Hidden Markov Models on a Smartphone with 3d Accelerometer, in: E. Corchado, M. Kurzyński, M. Woźniak (Eds.), Hybrid Artificial Intelligent Systems, Lecture Notes in Computer Science, 6678, Springer Berlin Heidelberg, 2011, pp. 460–467.

[12] E. Mitchell, D. Monaghan, N.E. O'Connor, Classification of sporting activities using smartphone accelerometers, Sensors 13 (4) (2013) 5317–5337.

[13] Y.-S. Lee, S.-B. Cho, Layered hidden Markov models to recognize activity with

built-in sensors on Android smartphone, Pattern Anal. Appl. (2016) 1–13, doi:10.1007/s10044-016-0549-8.

[14] P. Rashidi, A. Mihailidis, A survey on ambient-assisted living tools for older adults, IEEE J. Biomed. Health Inf. 17 (3) (2013) 579–590.

[15] E. Frontoni, P. Raspa, A. Mancini, P. Zingaretti, V. Placidi, Customers activity recognition in intelligent retail environments, in: International Conference on Image Analysis and Processing, Springer, 2013, pp. 509–516.

[16] S. Vishwakarma, A. Agrawal, A survey on activity recognition and behavior understanding in video surveillance, Vis. Comput. 29 (10) (2013) 983–1009.

[17] C. Xu, D. Tao, C. Xu, A survey on multi-view learning, arXiv preprint arXiv:1304.5634 (2013).

[18] D.H. Wolpert, Stacked generalization, Neural Netw. 5 (2) (1992) 241–259.

[19] A. Bobick, J. Davis, The recognition of human movement using temporal templates, Pattern Anal. Mach. Intell. IEEE Trans. 23 (3) (2001) 257–267, doi:10.1109/34.910878.

[20] P.C. Ribeiro, J. Santos-victor, Human activity recognition from video: modeling, feature selection and classification architecture, International Workshop on Human Activity Recognition and Modeling (HAREM), 2005.

[21] N. Robertson, I. Reid, A general method for human activity recognition in video, Comput. Vis. Image Understanding 104 (2) (2006) 232–248, doi:10.1016/j.cviu.2006.07.006.

[22] O.D. Lara, M.A. Labrador, A survey on human activity recognition using wearable sensors, Commun. Surv. Tutorials, IEEE 15 (3) (2013) 1192–1209.

[23] T. Brezmes, J.-L. Gorricho, J. Cotrina, Activity recognition from accelerometer data on a mobile phone, in: S. Omatu, M. Rocha, J. Bravo, F. Fernández, E. Corchado, A. Bustillo, J. Corchado (Eds.), Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living, Lecture Notes in Computer Science, 5518, Springer Berlin / Heidelberg, 2009, pp. 796–799.

[24] B. Khaleghi, A. Khamis, F.O. Karray, S.N. Razavi, Multisensor data fusion: a review of the state-of-the-art, Inf. Fusion 14 (1) (2013) 28–44.

[25] P.K. Atrey, M.A. Hossain, A. El Saddik, M.S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, Multimedia Syst. 16 (6) (2010) 345–379.

[26] A.P. James, B.V. Dasarathy, Medical image fusion: a survey of the state of the art, Inf. Fusion 19 (2014) 4–19. http://dx.doi.org/10.1016/j.inffus.2013.12.002.

[27] A. Tolstikov, X. Hong, J. Biswas, C. Nugent, L. Chen, G. Parente, Comparison of fusion methods based on DST and DBN in human activity recognition, J. Control Theory Appl. 9 (1) (2011) 18–27, doi:10.1007/s11768-011-0260-7.

[28] M. Amoretti, S. Copelli, F. Wientapper, F. Furfari, S. Lenzi, S. Chessa, Sensor data fusion for activity monitoring in the PERSONA ambient assisted living project, J. Ambient Intell. Humaniz. Comput. 4 (1) (2013) 67–84, doi:10.1007/s12652-011-0095-6.

[29] T. Hayashi, M. Nishida, N. Kitaoka, K. Takeda, Daily activity recognition based on DNN using environmental sound and acceleration signals, in: Signal Processing Conference (EUSIPCO), 2015 23rd European, 2015, pp. 2306–2310, doi:10.1109/EUSIPCO.2015.7362796.

[30] C. Zhu, W. Sheng, Human daily activity recognition in robot-assisted living using multi-sensor fusion, in: Robotics and Automation, 2009. ICRA'09. IEEE International Conference on, IEEE, 2009, pp. 2154–2159.

[31] O. Banos, M. Damas, A. Guillen, L.-J. Herrera, H. Pomares, I. Rojas, C. Villalonga, Multi-sensor fusion based on asymmetric decision weighting for robust activity recognition, Neural Process. Lett. 42 (1) (2015) 5–26.

[32] M. Nishida, N. Kitaoka, K. Takeda, Development and preliminary analysis of sensor signal database of continuous daily living activity over the long term, in: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific, IEEE, 2014, pp. 1–6.

[33] S. Sun, A survey of multi-view machine learning, Neural Comput. Appl. 23 (7) (2013) 2031–2038, doi:10.1007/s00521-013-1362-6.

[34] J. Zhao, X. Xie, X. Xu, S. Sun, Multi-view learning overview: recent progress and new challenges, Inf. Fusion 38 (2017) 43–54.

[35] A. Blum, T. Mitchell, Combining Labeled and Unlabeled Data with Co-training, in: Proceedings of the Eleventh Annual Conference on Computational Learning Theory, in: COLT' 98, ACM, New York, NY, USA, 1998, pp. 92–100, doi:10.1145/279943.279962.

[36] O. Chapelle, B. Schölkopf, A. Zien, others, Semi-Supervised Learning, MIT press Cambridge, 2006.

[37] Z.-H. Zhou, M. Li, Tri-training: exploiting unlabeled data using three classifiers, Knowl. Data Eng. IEEE Trans. 17 (11) (2005) 1529–1541.

[38] J. Wang, S.-w. Luo, X.-h. Zeng, A random subspace method for co-training, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, doi:10.1109/IJCNN.2008.4633789.

[39] Y. Yaslan, Z. Cataltepe, Co-training with relevant random subspaces, Neurocomputing 73 (10) (2010) 1652–1661.

[40] J. Farquhar, D. Hardoon, H. Meng, J.S. Shawe-taylor, S. Szedmak, Two view learning: SVM-2k, theory and practice, in: Advances in Neural Information Processing Systems, 2005, pp. 355–362.

[41] T. Diethe, D.R. Hardoon, J. Shawe-taylor, Multiview fisher discriminant analysis, In NIPS Workshop on Learning from Multiple Sources, 2008.

[42] Q. Chen, S. Sun, Hierarchical multi-view fisher discriminant analysis, in: International Conference on Neural Information Processing, Springer, 2009, pp. 289–298.

[43] Q. Wang, H. Lv, J. Yue, E. Mitchell, Supervised multiview learning based on simultaneous learning of multiview intact and single view classifier, Neural Comput. Appl. (2016) 1–9.

[44] M. Kubat, An Introduction to Machine Learning, Springer International Publishing, 2016.

[45] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, CRC press, 2012.

[46] L. Breiman, Stacked regressions, Mach. Learn. 24 (1) (1996) 49–64.

[47] P. Smyth, D. Wolpert, Linearly combining density estimators via stacking, Mach. Learn. 36 (1–2) (1999) 59–83.

[48] K.M. Ting, I.H. Witten, Issues in stacked generalization, J. Artif. Intell. Res.(JAIR) 10 (1999) 271–289.

[49] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[50] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. 15 (1) (2014) 3133–3181.

[51] P. Casale, O. Pujol, P. Radeva, Human Activity Recognition from Accelerometer Data Using a Wearable Device, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 289–296. 10.1007/978-3-642-21257-4_36

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

[53] L.T. Nguyen, M. Zeng, P. Tague, J. Zhang, Recognizing new activities with limited training data, in: Proceedings of the 2015 ACM International Symposium on Wearable Computers, ACM, 2015, pp. 67–74.

[54] C.E. Galván-Tejada, J.I. Galván-Tejada, J.M. Celaya-Padilla, J.R. Delgado Contreras, R. Magallanes-Quintanar, M.L. Martinez-Fierro, I. Garza-Veloz, Y. López-Hernández, H. Gamboa-Rosales, An analysis of audio features to develop a human activity recognition model using genetic algorithms, random forests, and neural networks, Mobile Inf. Syst. 2016 (2016) 1–10.

[55] I. Mendialdua, A. Arruti, E. Jauregi, E. Lazkano, B. Sierra, Classifier subset selection to construct multi-classifiers by means of estimation of distribution algorithms, Neurocomputing 157 (2015) 46–60.

[56] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, I. Rojas, Window size impact in human activity recognition, Sensors 14 (4) (2014) 6474–6499, doi:10.3390/s140406474.

[57] M. Zhang, A.A. Sawchuk, Motion primitive-based human activity recognition using a bag-of-features approach, in: ACM SIGHIT International Health Informatics Symposium (IHI), Miami, Florida, USA, 2012, pp. 631–640.

[58] E. Garcia-Ceja, R. Brena, Building personalized activity recognition models with scarce labeled data based on class similarities, in: J.M. García-Chamizo, G. Fortino, S.F. Ochoa (Eds.), Ubiquitous Computing and Ambient Intelligence. Sensing, Processing, and Using Environmental Information, Lecture Notes in Computer Science, 9454, Springer International Publishing, 2015, pp. 265–276, doi:10.1007/978-3-319-26401-1_25.

[59] M. Al Masum Shaikh, M. Molla, K. Hirose, Automatic Life-Logging: A novel approach to sense real-world activities by environmental sound cues and common sense, in: Computer and Information Technology, 2008. ICCIT 2008. 11th International Conference on, 2008, pp. 294–299, doi:10.1109/ICCITECHN.2008.4803018.

[60] U. Ligges, S. Krey, O. Mersmann, S. Schnackenberg, tuneR: Analysis of music, 2014.

[61] J.C. Gower, Some distance properties of latent root and vector methods used in multivariate analysis, Biometrika 53 (3–4) (1966) 325–338.

[62] A. Jalal, S. Kamal, D. Kim, Human depth sensors-based activity recognition using spatiotemporal features and hidden markov model for smart environments, J. Comput. Netw. Commun. 2016 (2016) 5, doi:10.1155/2016/8087545.

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

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

[65] C. Chen, R. Jafari, N. Kehtarnavaz, A real-time human action recognition system using depth and inertial sensor fusion, IEEE Sens. J. 16 (3) (2016) 773–781.

[66] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Forster, G. Troster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, J. del R Millan, Collecting complex activity datasets in highly rich networked sensor environments, in: Networked Sensing Systems (INSS), 2010 Seventh International Conference on, 2010, pp. 233–240, doi:10.1109/INSS.2010.5573462.

[67] Opportunity dataset, 2010, (https://archive.ics.uci.edu/ml/datasets/OPPORTUNITY+Activity+Recognition). Accessed: 30 March 2017.

[68] H. Sagha, S.T. Digumarti, J.d.R. Milln, R. Chavarriaga, A. Calatroni, D. Roggen, G. Trster, Benchmarking classification techniques using the opportunity human activity dataset, in: Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on, IEEE, 2011, pp. 36–40.