The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition

Ricardo Chavarriaga[a,*], Hesam Sagha[a], Alberto Calatroni[b], Sundara Tejaswi Digumarti[a], Gerhard Tröster[b], José del R. Millán[a], Daniel Roggen[b]

[a]*Chair on Non-Invasive Brain-Machine Interface, Center for Neuroprosthetics, School of Engineering, École Polytechnique Fédérale de Lausanne, Station 11, 1015 Lausanne, Switzerland*
[b]*Wearable Computing Laboratory, ETH Zürich, Gloriastrasse 35, 8092 Zürich, Switzerland*

Abstract

There is a growing interest on using ambient and wearable sensors for human activity recognition, fostered by several application domains and wider availability of sensing technologies. This has triggered increasing attention on the development of robust machine learning techniques that exploits multimodal sensor setups. However, unlike other applications, there are no established benchmarking problems for this field. As a matter of fact, methods are usually tested on custom datasets acquired in very specific experimental setups. Furthermore, data is seldom shared between different groups. Our goal is to address this issue by introducing a versatile human activity dataset recorded in a sensor-rich environment. This database was the basis of an open challenge on activity recognition. We report here the outcome of this challenge, as well as baseline performance using different classification techniques. We expect this benchmarking database will motivate other researchers to replicate and outperform the presented results, thus contributing to further advances in the state-of-the-art of activity recognition methods.

1. Introduction

Multiple applications require human activity recognition systems ranging from health care and assistive technologies (Tentori and Favela, 2008) to manufacturing (Stiefmeier et al., 2008) or gaming. New sensing technology allows the use of multimodal setups involving on-body, object-placed or ambient sensors. From a machine learning perspective activity recognition is a challenging problem as it typically deals with high-dimensional, multimodal streams of data characterised by a large variability (e.g. due to changes in the user's behaviour or as a result of noise). Moreover, real-life deployments are required to detect when no relevant action is performed (i.e. *Null* class) (Stiefmeier et al., 2008). For these reason robust methods are required tackling issues ranging from the feature selection and classification (Preece et al., 2009), to decision fusion and fault-tolerance (e.g., Chavarriaga et al. (2012); Sagha et al. (2011b); Zappi et al. (2007)).

However, the comparison of different approaches is often not possible due to the lack of common benchmarking tools and datasets that allow for replicable and fair testing procedures across several research groups. Currently, each research group assess the performance of their algorithms using experimental setups specially conceived for a narrow purpose. This contrasts with other application fields where publicly available datasets allow the independent assessment of different algorithms in the very same conditions. This is common practice in the machine learning community covering applications like computer vision, biometrics or speech recognition (e.g., the UCI machine learning repository). Furthermore, methods are often evaluated in the frame of open competitions or challenges providing a fair comparison of them. Recent examples of these competitions have focused on computer vision, bioinformatics, or brain-computer interfaces (e.g., Everingham et al. (2010); Guyon and Athitsos (2011); Blankertz et al. (2006)).

Considering this, we believe that there is a need for publicly available databases on human activity recognition. This will allow the replication of the testing procedures for different approaches. Ideally, these databases should reflect the variability of real-world activities, and be flexible enough to emulate different experimental setups and recording modalities. In order to address these issues we recorded a large recording of realistic daily life activities in a sensor-rich environment, i.e. The *Opportunity* activity recognition dataset (Roggen et al., 2010).

We used a subset of this dataset to organise an open challenge where different classification methods contributed by different research groups were compared. The selected benchmarking dataset is publicly available and contains recordings of on-body sensors while subjects perform activities of daily living, ranging from simple motion primitives to complex gestures. Thus, this dataset offers a rich playground to assess methods for sensor selection, feature extraction, classifier calibration and adaptation, multimodal data fusion, automatic segmentation, among others. It also captures the challenges common to many other activity recognition scenarios. Thus, methods proved to be robust on this dataset can likely be successfully translated to other activity recognition problems.

*Corresponding author: Ricardo Chavarriaga
Email address: ricardo.chavarriaga@epfl.ch
(Ricardo Chavarriaga)

This paper provides an overview of the Opportunity dataset (Section 3) and the activity recognition challenge (Sections 4, 5). Furthermore, we also compare different recognition systems using four well-known classification techniques, namely k–NN, NCC, LDA and QDA classifiers (Section 6). The performance of these methods and the contributed techniques are then reported (Section 7), followed by the conclusions (Section 8).

2. Related work

Several datasets for activity recognition are currently available. However, they tend to be specific to an activity recognition purpose. Widely popular in the pervasive computing community is the *PlaceLab* dataset. It contains ambient and object sensing of subjects recorded over several days (up to a week) in an environment with multimodal sensors (Intille et al., 2006). Its main strength is to provide long-term recordings although it does not include a high number of activity instances. Another dataset recorded by van Kasteren et al. (2008) features longer recordings (month-long) but fewer sensors. It uses digital or binary sensors (e.g. reed switches) to record interactions with objects of interest, but does not include information about modes of locomotion or body posture. The *Darmstadt routine dataset* –used to study unsupervised activity pattern discovery (Huynh et al., 2008)– is a long recording from body activity collected by the Porcupine system (Van Laerhoven et al., 2006). The *TUM Kitchen dataset* focuses on video-based activity recognition (Tenorth et al., 2009), and also contains RFID and reed switch data, but it does not include on-body sensors. A more recent database focuses on fine grained human activities in the kitchen, but it is more suitable for computer vision techniques (Rohrbach et al., 2012). As it can be seen, these databases –although useful– are limited due to the reduced number of recorded sensors and activity instances as well as the fact that they were conceived for very specific purposes.

An exceptional effort to collect a large scale human activity corpus, termed *HASC corpus*, was led by Nagoya University (Kawaguchi et al., 2011). It is the result of a collaboration among 20 teams that gathered data from 116 subjects. Data from each subject contains a set of six activities (stay, walk, jogging, skip, stair-up and stair-down) recorded with a single commercially available accelerometer. There was no constraint on the location of the sensor. The main strength of this corpus lies in the large amount of subjects available. However, the fact that there is only one sensor that may be located at any place, effectively limits its use.

In addition, most of the previous activity recognition challenges have focused on isolated gestures using video (Guyon and Athitsos, 2011). One exception to this is the open contest organised at the 2011 Body sensor network conference (http://bsncontest.org) (Giuberti and Ferrari, 2011). In this contest, the organisers provided three datasets provided from different groups. Datasets differ in the number, arrangement and type of sensors used, as well as the number of subjects. Participants were asked to provide methods for the recognition of several actions mainly focusing on modes of locomotion.

The lack of more general databases can be explained by the difficulty to conceive and record a dataset that reflects the complexity and variability of daily life situations. Moreover, proper comparison of machine learning techniques requires these datasets to provide a reasonable amount of instances for the different recorded actions and to include several subjects in order to allow the assessment of inter-subject variability. In addition, if the database is used to emulate changes in the sensor network, then activities should be recorded by a large and diverse set of sensors. These aspects were taken into account for the database described in the next section.

3. The Opportunity dataset

The challenge was based on a subset of the *Opportunity activity recognition dataset* (Roggen et al., 2010), a dataset of complex naturalistic activities with a particularly large number of atomic activities (more than 27'000) collected in a sensor rich environment (c.f. Figure 1). Overall, it comprises recordings of 12 subjects using 15 networked sensor systems, with 72 sensors of 10 modalities, integrated in the environment, in objects, and on the body (Figure 1b). These characteristics make it well suited to benchmark various activity recognition approaches (Sagha et al., 2011a). An illustrative video of the recording and database is provided as supplementary material.

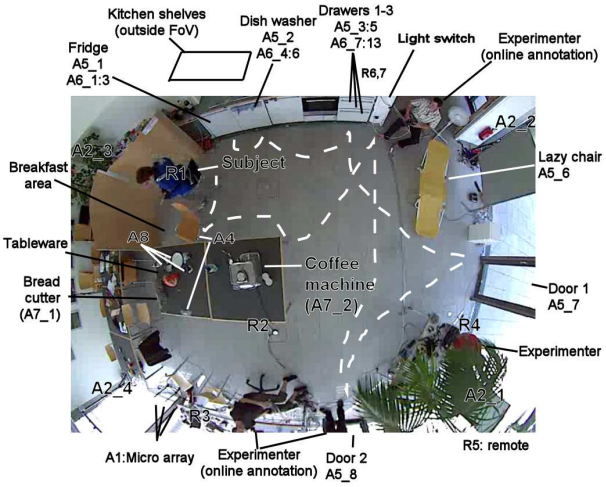
We designed the activity recognition environment and scenario to generate many activity primitives, yet in a realistic manner. We purposely did not record human behaviour in daily life to favour the use of a highly multimodal setup. Instead, we aimed at maximising the number of activity instances collected, while keeping their execution naturalistic. We achieved this by relying on a high-level script and leaving free interpretation to the users, and even encouraging them to perform as naturally as possible with all the variations they were used to. Subjects operated in a room simulating a studio flat with a deckchair, a kitchen, doors giving access to the outside, a coffee machine, a table and a chair.

3.1. Scenario script

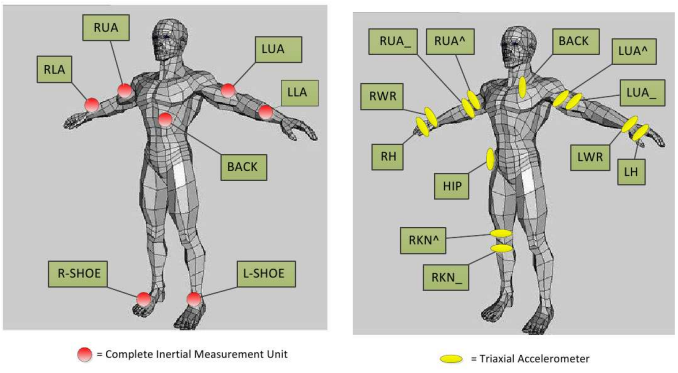
For each subject we recorded 6 different runs. Five of them, termed activity of daily living (ADL), followed a given scenario as detailed below. The remaining run (termed *drill* run) was designed to collect a large number of activity instances. The ADL run consists of temporally loosely defined situations. Each situation (e.g. preparing sandwich) was annotated in terms of composite activities (e.g. cutting bread) as well as atomic activities (e.g. reach for bread, move to bread cutter, operate bread cutter). Thus allowing to analyze the data at activity recognition at various abstraction levels (Roggen et al., 2010).

3.1.1. ADL runs

These runs consist of daily morning activities. The subject starts lying on the deckchair, then s/he gets up and move around in the kitchen checking the objects in drawers and shelves, before walking outside the room. When s/he returns, the subject prepares a coffee with milk and sugar. After drinking the coffee, s/he prepares and eats a sandwich with salami and cheese.



(a)



(b)

Figure 1: Opportunity dataset setup. (a) Top view of the recording room. The dashed line shows a typical user trajectory in the drill run. (b) On-body sensors used for the activity recognition challenge (Red: IMU sensors; Yellow: 3-axis accelerometers).

Finally s/he should clean the room, putting the objects in their original location or in the dishwasher before going back to the deck chair. It should be noticed that there is no constrain on the location or body posture in any of the scripted activities.

3.1.2. Drill run

This run is intended to record a large set of activity instances. To achieve so, subjects performed 20 repetitions of the following sequence :

1. Open and close the fridge
2. Open and close the dishwasher
3. Open and close 3 drawers (at different heights)
4. Open and close door 1
5. Open and close door 2
6. Turn on and off the lights
7. Clean table
8. Drink (standing)
9. Drink (sitting)

3.2. Sensor systems

The deployed sensors include 24 custom Bluetooth wireless accelerometers and gyroscopes, 2 Sun SPOTs and 2 InertiaCube3, the Ubisense localisation system and a custom-made magnetic field sensor (Pirkl et al., 2008). Seven computers acquired the data from specific sensor systems. On-body sensors were managed by a dedicated laptop in a backpack (local storage as there was no WLAN outside of the room). Ambient and object sensors were acquired by multiple computers according to the bandwidth required, cabling possibilities, and the need to minimise the risk of data loss. This type of recording requires special attention to the synchronisation of multiple streams, curation and annotation of a large amount of raw data, as well as development of appropriate tools (Calatroni et al., 2011).

4. Challenge on Robust Activity Recognition

For the activity recognition challenge, we use a subset of the Opportunity database corresponding to recordings of on-body sensors for 4 subjects. These sensors include five XSense inertial measurement units (accelerometer, gyro and magnetic sensors) mounted on a custom-made motion jacket, 12 Bluetooth 3-axis acceleration sensors on the limbs and commercial InertiaCube3 inertial sensors located on each foot (Figure 1b). The challenge ran from May to September 2011 and the outcome was initially announced during a workshop at the IEEE conference on Systems, Man and Cybernetics in Anchorage, Alaska (Sagha et al., 2011a).

We define four different tasks targetting the recognition of modes of locomotion and arm gestures (Task A and B2, respectively); activity spotting (Task B1) and robustness to noise (Task C). The activities selected for the challenge are summarised in Table 1. Since subjects had a lot of freedom when performing the ADL scenario, there is a large variability in the length and number of instances of the different classes, especially in the case of gestures.

The data was made publicly available and the annotated labels of selected sessions were provided for training purposes. Labelled sessions include the full recording of Subject 1, as well as for three ADL sessions of subjects 2, 3 and 4. For the evaluation of the contributed methods the labels of sessions ADL4 and ADL5 of subjects 2,3 and 4 were withhold until the end of the challenge. We use the Weighted F-measure (c.f. Section 5.1) of all activity classes to rank the contributed methods.

In order to evaluate the robustness of the recognition methods, rotational noise was added to the testing sessions of subject 4 (See section 4.1.4). This type of noise can affect body worn sensors, e.g. as a result of placement differences across or within sessions. The rotational noise (up to 60°) was added at random times for each IMU, affecting all the sensors in the IMU (accelerometer, gyro, and magnetic sensors).

4.1. Challenge tasks

The challenge is composed of four tasks addressing different aspects of the activity recognition problem. These tasks are:

Table 1: Class labels for both modes of locomotion (Task A) and gesture recognition (Tasks B1,B2, C). The numbers in parentheses denote the number of instances recorded during the ADL runs (all subjects combined).

Modes of locomotion	Gestures			
Stand (1093)	open Dishwasher (50)	open Drawer1 (50)	open Drawer2 (44)	open Drawer3 (56)
Sit (1095)	close Dishwasher (56)	close Drawer1 (49)	close Drawer2 (44)	close Drawer3 (57)
Walk (90)	open Fridge (129)	open Door1 (45)	open Door2 (43)	move Cup (184)
Lie (40)	close Fridge (133)	close Door1 (39)	close Door2 (41)	clean Table (33)
Null	*Null*			

4.1.1. Task A: Multimodal activity recognition: Modes of locomotion

The goal of this task is to classify modes of locomotion from the full set of body-worn sensors (c.f. Table 1). The testing dataset for this task is composed of data from Subjects 2 and 3 (ADL4, ADL5). This is a 4-class continuous activity recognition problem.

4.1.2. Task B1: Automatic segmentation

Typically, activity recognition methods are evaluated using recordings that have already been segmented into the different target classes. However, realistic deployments are required to detect when no relevant action is performed (i.e. *Null* class). This involves the detection of the specific time when relevant actions begin and end within a continuous recording.

This is a 2-class segmentation problem, (*Null* vs. activity class). The activity class comprises all gestures (c.f. Table 1), and binary labels denote whether any of them is being executed.

The full set of sensors is considered for this task (i.e. the motion jacket, 12 bluetooth body-worn accelerometers and inertial sensors on the feet). The testing dataset for this task is composed of data from Subjects 2 and 3 (ADL4, ADL5).

4.1.3. Task B2: Multimodal activity recognition: Gestures

This task concerns recognition of the different right-arm gestures, as described above. We provided unsegmented labelled data sets for the gestures listed in Table 1.

This is a 17 class segmentation and classification problem. As for the previous cases the full set of sensors are considered for this task and the same testing . dataset is used. (ADL4, ADL5 of Subjects 2 and 3).

4.1.4. Task C: Robustness to noise: Gestures

Realistic applications are prone to noise due to different factors. This task focuses on methods that are robust to sensor noise. As explained above, for this task rotational noise has been added to the testing dataset. The gestures to be recognised are the same as for Task B2, but only the motion jacket sensors are considered. The testing dataset for this task is composed of data from Subject 4 (ADL4, ADL5)

5. Performance Measures

Choosing an appropriate measure to assess the performance of complex activity recognition systems is not straightforward. Some measures may reflect specific qualities of the system while hiding or misrepresenting others (Ward et al., 2011). In particular, dealing with continuous real-life data adds new dimensions to this problem. On the one hand, labels used as ground truth might be loosely defined or ambiguous (i.e. the time when a gesture starts or finishes is subjectively assessed by the person doing the labelling). On the other hand, these recordings contain periods of time when none of the predefined class actions is performed. However, in these periods it cannot be assumed that the person remained still; most of the time s/he is performing another action or is in a transition from one action to another. Moreover, continuous recordings often lead to highly unbalanced datasets with some classes being overrepresented with respect to the others. For instance, in the gesture recognition case the *Null* class represents more than 75% of the recorded data (76%, 82%, 76% and 78% for subjects 1 to 4, respectively). Taking these points into consideration, we discuss and present different types of performance measures computed in the challenge dataset.

5.1. Weighted F - Measure

The F-measure (F_1) takes into account the precision and recall for each class and can provide a better assessment of performance than computing the accuracy (correct predicted/number of samples). Precision is defined as $\frac{TP}{TP+FP}$, and recall corresponds to $\frac{TP}{TP+FN}$; where TP, FP are the number of true and false positives, respectively, and FN corresponds to the number of false negatives.

Furthermore, to counter the class imbalance, classes can be weighted according to their sample proportion,

$$F_1 = \sum_i 2 * w_i \frac{precision_i.recall_i}{precision_i + recall_i} \quad (1)$$

where i is the class index and $w_i = n_i/N$ is the proportion of samples of class i, with n_i being the number of samples of the i^{th} class and N being the total number of samples.

5.2. Area under the ROC curve

We also report the area under the curve (AUC) in the ROC (Receiver Operating Characteristic) space. This measure takes into account the sensitivity (recall) and specificity of the classification. The ROC curve is a plot between sensitivity($\frac{TP}{TP+FN}$) and specificity ($\frac{TN}{FP+TN}$, TN: True negatives) drawn for different threshold values for each class.

As with the F-measure, the class imbalance can be taken into account by weighting the AUC for each class by its prevalence on the data (Fawcett, 2006).

$$AUC_{total} = \sum_i w_i AUC(c_i) \quad (2)$$

where AUC_{total} reflects the overall performance, w_i is the weight for the i^{th} class, and $AUC(c_i)$ is the AUC for the i^{th} class.

5.3. Misalignment Measures

Although the two methods presented above generally give a good assessment of the classification performance, their results may be misleading when recognising actions from continuously recorded data. Indeed, as the onset and offset times of an action are not precisely defined, misalignment of the output labels (e.g. early detection of an action onset) may be wrongly considered as classification errors. This becomes more critical as annotations are performed by human operators, that may have different criteria for defining the actual onset of an action. In order to address this issue, Ward et al. (2011) explicitly defined different types of errors as follows,

Overfill: When the start (or stop) time of a predicted label is earlier (or later) than the actual time.

Underfill: When the start (or stop) time of a predicted label is later (or earlier) than the actual time.

Insertion: Predicting an action when there is no activity of interest (i.e. *Null* class).

Merge: When subsequent actions of the class are recognized as a single one (i.e. ignoring the *Null* class between them).

Fragmentation: Predicting *Null* class in between an uninterrupted activity class.

Deletion: Assigning a *Null* label when there is an activity.

Substitution: When an activity is misclassified as a different class (other than *Null*).

As noted above, label misalignment may result on overfill or underfill errors. Clearly, the impact of this type of errors, as opposed to deletion or substitution, will be application dependent.

6. Baseline performance

As a baseline for evaluating different recognition methods, including those submitted by challenge participants, we report the performance of commonly used classification methods on the tasks proposed in the challenge. Following the challenge guidelines, only on-body sensors were taken into account; i.e. Five IMUs on the motion jacket, 2 InertiaCube3 sensors in the feet and 12 bluetooth 3-axis acceleration sensors (c.f. Section 4; Figure 1b). Each sensor axis is treated separately yielding an input space dimensionality of 113 for tasks A,B1, and B2[1] and 45 for task C, where only the motion jacket is considered.

We train user-specific classifiers, always using as training set the data of ADL1, ADL2, ADL3 and Drill sessions. We report classification performance for each subject on a testing set composed of ADL4 and ADL5. We emphasise that the goal of this analysis is to provide a reference performance instead of obtaining a highly accurate recognition system. For this reason we chose simple, well known processing and classification methods, which may also be the most likely to be included at first in upcoming commercial activity-aware products.

There is a considerable amount of missing data in the dataset mainly due to disconnection of wireless sensors. Although more complex methods have been proposed to tackle this issue (c.f. Saar-Tsechansky and Provost (2007)), in this study we simply repeat the previous value in place of the missing value.

Since the data was recorded continuously and it is not segmented, we used a sliding window of length 500ms with a step size of 250ms for extracting the features. We used as feature the mean value of the sensor on each window.

We tested four well known classification techniques: k–Nearest Neighbours (k–NN), Nearest Centroid Classifier (NCC) and two Gaussian classifiers, Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), as described below.

- *k–Nearest Neighbours (k–NN) classifier*: This is the simplest classifier used where the Euclidean distances between a test sample and the samples of the training set are computed and the most frequently occurring label of 'k'-closest samples is selected as output label. In our analysis we used k–NN with two values of 'k', 1 and 3. For this classifier all the training samples are stored and the *Null* class is explicitly modelled.

- *Nearest Centroid Classifier (NCC)*: In this method, the Euclidean distance between the test sample and the centroid for each class samples is used for the classification. We first train the classifier only on those samples that correspond to activities of interest excluding *Null* samples. Then using the entire training set we estimate class specific thresholds that maximise the F-measure for each class.

- *Linear Discriminant Analysis (LDA)*: This is a Gaussian Classifier which classifies on the assumption that the features are normally distributed and all classes have the same covariance matrix. For detection of the *Null* class, we used the same rejection mechanism as for NCC.

- *Quadratic Discriminant Analysis (QDA)*: Similar to the LDA this technique also assumes a normal distribution for the features but the class covariances may differ. The same rejection mechanism as NCC and LDA is used.

In the following section we present the classification results using different measures for these classifiers, as well as the contributions submitted to the challenge.

7. Results

7.1. Challenge submissions

The challenge was publicly announced on the week of the 20th of may 2011 on several mailing lists (including ubicomp,

[1] The feature vector comprises 9 dimensions per IMU on the jacket, 16 values per InertiaCube sensors, and 3 per bluetooth accelerometer.

Table 2: Summary of challenge submissions

Code	Institution	Sensors and Features	Missing data	Classifier	Tasks
UP	U. of Parma, Italy	Mean, std, min, max, duration$^{N/A}$	-	Comparison	A,B1,B2,C
NStar	A*Star, Singapore	Normalized values^{r}	Spline	1-NN^{m}	A,B1,B2
SStar	A*Star, Singapore	Normalized values^{r}	Linear	SVM^{m,s}	A,B1,B2
CStar	A*Star, Singapore	Normalized values^{r}	Spline	SVM + 1-NN^{m,s}	A,B1,B2
NU	Nagoya U., Japan	Mean, var, energy†	-	C4.5^{w}	A
MI	Masdar Inst, Abu Dhabi	PCA on sensor values^{a}	-	k-NN^{w}	A
MU	Monash U., Australia	Sensor values^{a}	-	DT grafting^{w}	A
UT	U. of Tokyo	PCA on value, mean and var‡	Skip	Adaboost	A
NAGS	IIT Bombay, India	Discretized values$^{N/A}$	-	HMM	C

Sensors and Features. *N/A*: No information available. *a*: All sensors. *r*: Removed RH accelerometer due to high number of missing values. †: Accelerometers RKN^, BACK, L-SHOE. ‡: Subject-specific manually selected sensors.

Missing data. Spline: Spline interpolation. Linear: Linear interpolation. Skip: Skip sample and repeat last decision.

Classification. DT grafting: Decision tree grafting (Webb, 1999). *w*: Weka implementaion. *m*: Matlab implementation.

Pascal network, Panorama project, etc.) as well as personal contacts made by the organisers. The announcement drawn more than 2'000 visits to the Challenge webpage, and about 700 downloads of the dataset in the span of 4 months.

There were nine contributions from seven teams as summarized in Table 2. Overall eight groups tackled the task A, while Task C was the one with less submissions (two submissions). Most contributions used the sensor values as features, sometimes normalized or discretized, while two contributions used PCA to reduce the dimensionality space. Regarding missing values, three methods use signal interpolation (linear or spline), while another group just repeated the decision of the last available sample. Different standard classification algorithm where used including Decision trees, k–NN, SVM and HMMs. Three of the contributions use the WEKA machine learning tool (Hall et al., 2009), two of them use LIBSVM library (Chang and Lin, 2011) and three contributions use Matlab implementations.

7.2. Classification performance

We report here the classification performance using methods described in Sections 5 and 6, as well as the challenge contributions. In the former case we report results on all 4 subjects. In the case of the challenge submissions we report the performances on subjects 2 and 3 (i.e. Tasks A and B2; modes of locomotion and gesture recognition, respectively), as well as subject 4 (i.e. Task C; noisy data).

7.2.1. F-measure and AUC

The weighted F-measure–either including or not the *Null* class[2]–is reported in Tables 3 and 4. In general, k–NN classifiers perform best for both locomotion and gestures recognition, followed by the Gaussian classifiers. As expected the class imbalance, in particular the inclusion of the *Null* class, may lead to overstimation of the performance for gesture recognition. The weighted AUC measure is consistent with the F-measures where both LDA and QDA outperform NCC (c.f. Table 5).

The performance of the contributed methods is also presented in Tables 3 and 4. Regarding the measure used in the challenge, F-measure of activity classes (without *Null* class), the mode of locomotion is recognized reasonably well with five of them performing above 0.85. Nevertheless, the best method was not largely better than the k–NN classifiers. When the *Null* class was considered, all contributions but one (MI, using k–NN classifier) decreased considerably their performance.

In contrast, the contributed methods for gesture recognition outperform the baseline classifiers when there is no rotational noise (Task B2). None of the two submissions for Task C perform better than the k–NN or Gaussian classifiers. As observed before, given the large number of *Null* class samples a large decrease in performance is observed when considering only the activity classes.

7.2.2. Misalignment Measures

The measures proposed by Ward et al. (2011) are shown in Figures 2 and 3 for the baseline classifiers. They follow the same pattern as the F-measures, yielding higher performance for the k–NN classifiers, even with noisy data. Regarding the recognition of modes of locomotion, it is worth noticing that these classifiers have a small rate of overfill and underfill errors, suggesting that they accurately capture the on/offset of the actions. However, this percentage increases for subject 4 when the rotational noise is added and only sensors on the upper torso are available.

In the case of gesture recognition, k–NN classifiers have a reduced level of insertions errors. This suggests that the *Null* class is not properly discriminated by the threshold-based rejection mechanism. Moreover the amount of underfill errors is lower for k–NN, although the number of overfill errors is similar to other methods.

[2]Note that this measure disregards the true negatives (correctly classified *Null*-class samples), while taking into account false negatives.

Table 3: Recognition of modes of locomotion (F-measure). Methods contributed to the challenge are presented in the bottom rows of the table. In all tables the column [S2 S3] corresponds to the concatenated data from subjects 2 and 3. The performance on these data was used to rank the challenge submissions. Boldface denote the method(s) with highest performance.

Modes of Locomotion - Task A

	F measure					F measure (No *Null* class)				
Method	S1	S2	S3	[S2 S3]	S4	S1	S2	S3	[S2 S3]	S4
LDA	0.62	0.64	0.68	0.59	0.43	0.73	0.70	0.74	0.64	0.53
QDA	0.67	0.66	0.71	0.68	0.45	0.81	0.77	0.79	0.77	0.56
NCC	0.60	0.58	0.56	0.54	0.41	0.69	0.67	0.62	0.60	0.50
1 NN	0.84	0.85	0.83	**0.84**	0.76	0.85	**0.85**	**0.85**	**0.85**	0.76
3 NN	**0.85**	**0.86**	**0.83**	**0.85**	**0.77**	**0.86**	**0.86**	**0.85**	**0.85**	**0.76**
UP		0.58	0.62	0.60			0.88	0.80	0.84	
NStar		0.58	0.66	0.61			0.88	0.85	0.86	
SStar		0.61	0.68	0.64			0.87	0.83	0.86	
CStar		0.60	0.65	0.63			**0.90**	0.83	0.87	
NU		0.54	0.49	0.53			0.83	0.63	0.75	
MI		**0.85**	**0.81**	**0.83**			0.87	0.86	0.86	
MU		0.57	0.68	0.62			0.86	**0.87**	**0.87**	
UT		0.48	0.55	0.52			0.74	0.72	0.73	

Table 4: Gesture recognition (F-measure). Methods contributed to the challenge are presented in the bottom rows of the table.

Gesture recognition - Tasks B2 ([S2 S3]), C(S4)

	F measure					F measure (No *Null* class)				
Method	S1	S2	S3	[S2 S3]	S4	S1	S2	S3	[S2 S3]	S4
LDA	0.65	0.63	0.70	0.69	0.62	0.36	0.28	0.27	0.25	0.17
QDA	0.60	0.57	0.69	0.53	0.64	0.34	0.29	0.34	0.24	0.22
NCC	0.48	0.48	0.51	0.51	0.35	0.29	0.21	0.22	0.19	0.14
1 NN	**0.85**	0.89	0.86	**0.87**	0.84	0.56	0.53	**0.58**	**0.55**	0.46
3 NN	0.85	**0.89**	**0.86**	0.85	**0.88**	**0.55**	**0.53**	**0.58**	**0.56**	**0.48**
NStar		0.84	0.83	0.84			0.60	0.69	0.65	
SStar		0.87	0.84	0.86			0.65	0.72	0.70	
CStar		**0.88**	**0.87**	**0.88**			**0.72**	**0.80**	**0.77**	
UP		0.64	0.64	0.64	0.64		0.23	0.19	0.22	0.16
NAGS					**0.71**					**0.17**

7.3. Effect of Null *rejection threshold*

As mentioned above we use a threshold-based mechanism to detect samples of the *Null* class (c.f. Section 6). The results reported in Section 7.2 use a class-specific threshold based on the training data. Here we test how much the performance is affected if a class-independent threshold is defined. In this case we set the common threshold as the mean value of the class-specific thresholds previously found. Unsurprisingly, a common threshold affects the performance, mainly decreasing the rate of true negatives (compare Figures 3 and 4a). In general, there is a decrease in the amount of true negatives in the classification of gestures, and an increase in the number of deletions, mostly for the Gaussian classifiers.

7.4. Effect of sensor choice and rotational noise

In the previous results the data from subject 4 include a smaller set of sensors and test data was noisy. To compare the effect of the rotational noise and the number of sensors, we performed the complete analysis for all subjects considering only the motion jacket sensors (no noise added). We observe a small decrease in performance in comparison to the results using all sensors (c.f. Figures 3 and 5a, respectively). Moreover, when comparing the performance of subject 4 with and without noise, we observe that performance decrease in the first case, specially for the Gaussian and NCC classifiers

8. Conclusion

This paper presents the outcome of the Opportunity activity recognition challenge. It illustrates the use of a common database to assess performance of different methods over several subjects and recording conditions. We study the recognition of modes of locomotion and gestures using data from 4 subjects performing daily activities recorded with different inertial sensor modalities. Moreover, one of the subjects has a different sensor configuration and noisy data.

As a baseline for future studies, we report the performance achieved by standard classification techniques such as k-NN,

Table 5: Classification performance in terms of area under the ROC curve (AUC).

	Modes of Locomotion					Gesture recognition				
Method	S1	S2	S3	[S2 S3]	S4	S1	S2	S3	[S2 S3]	S4
LDA	0.76	0.77	0.77	0.74	0.63	0.86	**0.76**	0.85	0.80	0.79
QDA	**0.84**	**0.82**	**0.84**	**0.82**	**0.67**	**0.87**	**0.76**	**0.89**	**0.82**	**0.90**
NCC	0.79	0.72	0.74	0.70	0.61	0.76	0.70	0.77	0.73	0.70

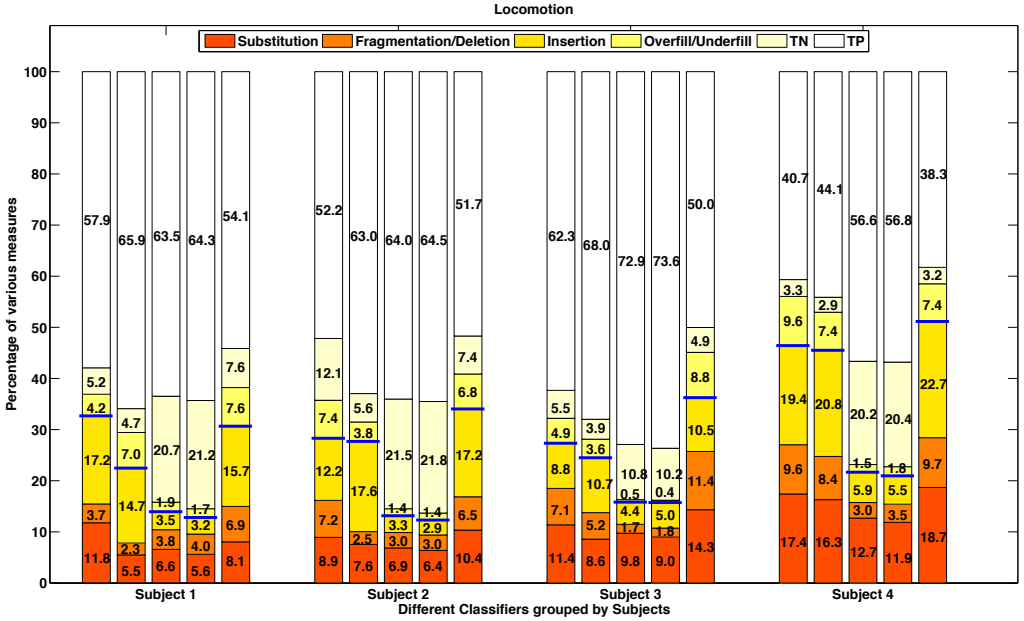


Figure 2: Modes of locomotion (Task A) - Performance evaluation on the Opportunity challenge dataset using the measures proposed by Ward et al. (2011). Each group of five columns denotes the accuracy of LDA, QDA, 1-NN, 3-NN and NCC, respectively. Note that the data of subject 4 has rotational noise added and only sensors on the motion jacket are used.

NCC, LDA, and QDA. These results highlight the effect of class imbalance in the computed F-measures and overall performance. As a matter of fact, in continuous, unsegmented data – as obtained in real-life conditions and provided in the database– the *Null* class is typically overrepresented. Furthermore, features of these class likely overlap with the activities of interest. This aspect should not be neglected at the design stages (e.g. by using risk functions when optimising the classifier parameters).

Methods contributed to the challenge were mostly based on well established classification techniques. For the classification of modes of locomotion, these methods perform at a similar level as the proposed baseline. In contrast, they perform better for gesture recognition. Only one of the proposed methods fuses different classification techniques, in this case 1-NN and SVM. Overall, this method outperforms the others for gesture segmentation and recognition (Task B1, B2). Especially when the Null class is not taken into account. We note that while a large number of sensors were provided, no work did use sensor or feature selection approaches. Similarly, only four submissions explicitly addressed the issue of missing values, either by interpolation or repetition of previous decisions. Among them, those who use interpolation consistently obtain high performance.

One main goal of this work is to promote data sharing and comparison of recognition methods using common benchmarks. To achieve this, the described dataset is made publicly available, including matlab scripts allowing to reproduce the results here presented. Data and scripts are available at the UCI Machine learning repository[3]. Moreover, the full Opportunity dataset, including wearable and ambient sensors recorded on 10 subjects is also available[4]. We hope this initiative will allow other researchers to replicate and outperform the presented results, thus assessing the improvement that can be achieved using more complex techniques.

Furthermore we stress that this rich dataset offers many opportunities for further comparative assessments and possibly new challenges. A few non-exhaustive examples of upcoming research activities that may benefit from this dataset include e.g:

Sensor and feature selection techniques to find the most suitable features to use for activity recognition, and whether raw sensor readings or composite information based on additional

[3] http://archive.ics.uci.edu/ml/datasets/OPPORTUNITY+Activity+Recognition
[4] http://www.contextdb.org

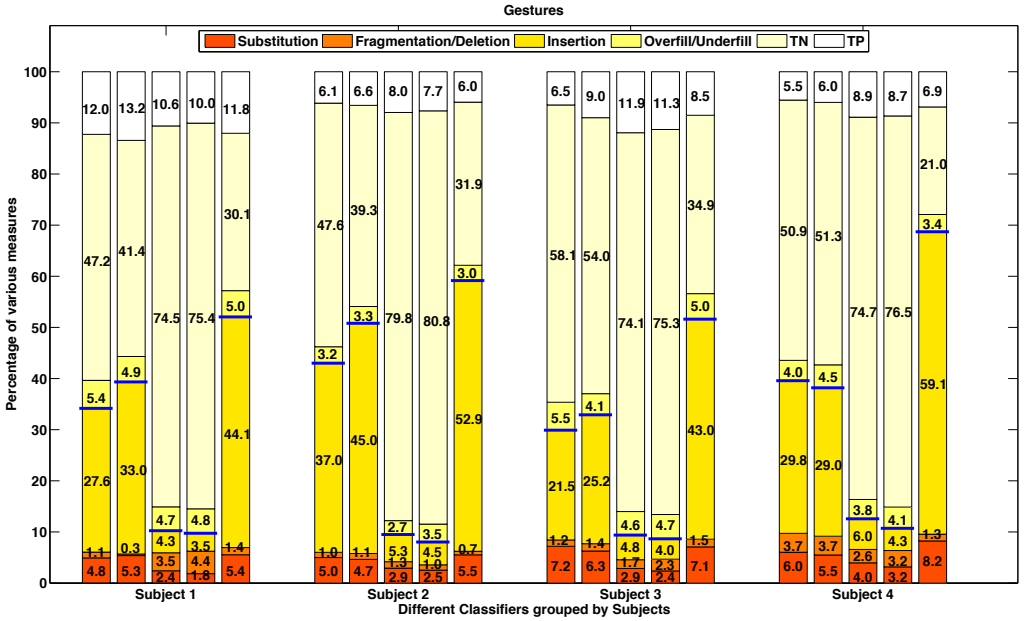


Figure 3: Gesture recognition (Tasks B2,C) - Performance evaluation on the Opportunity challenge dataset using the measures proposed by Ward et al. (2011). Each group of five columns denotes the accuracy of LDA, QDA, 1-NN, 3-NN and NCC, respectively.

knowledge is most adequate (e.g., Zinnen et al. (2009)).

Dynamic multimodal data fusion. In particular the dynamic selection of the best sensor configuration according to their availability, as well as runtime exploitation of new resources (see Banos et al. (2012) for an example on transfer learning techniques).

Hierarchical activity inference. Inferring high-level activities from low-level primitives opens the door to investigate the identification of relevant action primitives, and the combination of machine learning and reasoning techniques.

Finally, this dataset may be useful for assessing novel approaches including semi-supervised learning, active learning and hierarchical clustering relying on sparse labels or assisting experts by spotting relevant areas in the dataset.

In summary, the main characteristics of the dataset comprising a large number of sensors, the realism of the activities recorded, as well as their complexity and careful annotation at a fine-grained and high-level, allow its use for studying a wide range of aspects of activity recognition and provide a shared tool for successful method comparison across the community.

9. Acknowledgements

A preliminary version of this work was presented at the IEEE conference on Systems, Man and Cybernetics in Anchorage, Alaska (Sagha et al., 2011a). This work has been supported by the EU Future and Emerging Technologies (FET) contract number FP7-Opportunity-225938. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

References

Banos, O., Calatroni, A., Damas, M., Pomaras, H., Rojas, I., Sagha, H., Millán, J.d.R., Tröster, G., Chavarriaga, R., Roggen, D., 2012. Kinect=IMU? Learning MIMO Signal Mappings to Automatically Translate Activity Recognition Systems Across Sensor Modalities, in: International Symposium on Wearable Computers, pp. 92–99.

Blankertz, B., Müller, K.R., Krusienski, D., Schalk, G., Wolpaw, J.R., Schlögl, A., Pfurtscheller, G., Millán, J.d.R., Schröder, M., Birbaumer, N., 2006. The BCI competition III: Validating alternative approaches to actual BCI problems. IEEE Trans Neural Syst Rehab Eng 14, 153–159.

Calatroni, A., Roggen, D., Tröster, G., 2011. Collection and curation of a large reference dataset for activity recognition, in: Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on, pp. 30 –35.

Chang, C.C., Lin, C.J., 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27.

Chavarriaga, R., Bayati, H., Millán, J., 2012. Unsupervised adaptation for acceleration-based activity recognition: robustness to sensor displacement and rotation. Personal and Ubiquitous Computing 10.1007/s00779-011-0493-y.

Everingham, M., Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (VOC) challenge. Int. J. Comput. Vision 88, 303–338.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters 27, 861 – 874.

Giuberti, M., Ferrari, G., 2011. Simple and robust BSN-based activity classification: Winning the first BSN contest, in: Proc. 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL), Barcelona, Spain. pp. 34:1–34:5.

Guyon, I., Athitsos, V., 2011. Demonstrations and live evaluation for the gesture recognition challenge, in: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pp. 461 –462.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. SIGKDD Explor. Newsl. 11, 10–18.

Huynh, T., Fritz, M., Schiele, B., 2008. Discovery of activity patterns using topic models, in: Proceedings of the 10th international conference on Ubiquitous computing, ACM New York, NY, USA. pp. 10–19.

Intille, S., Larson, K., Tapia, E., Beaudin, J., Kaushik, P., Nawyn, J., Rockinson,

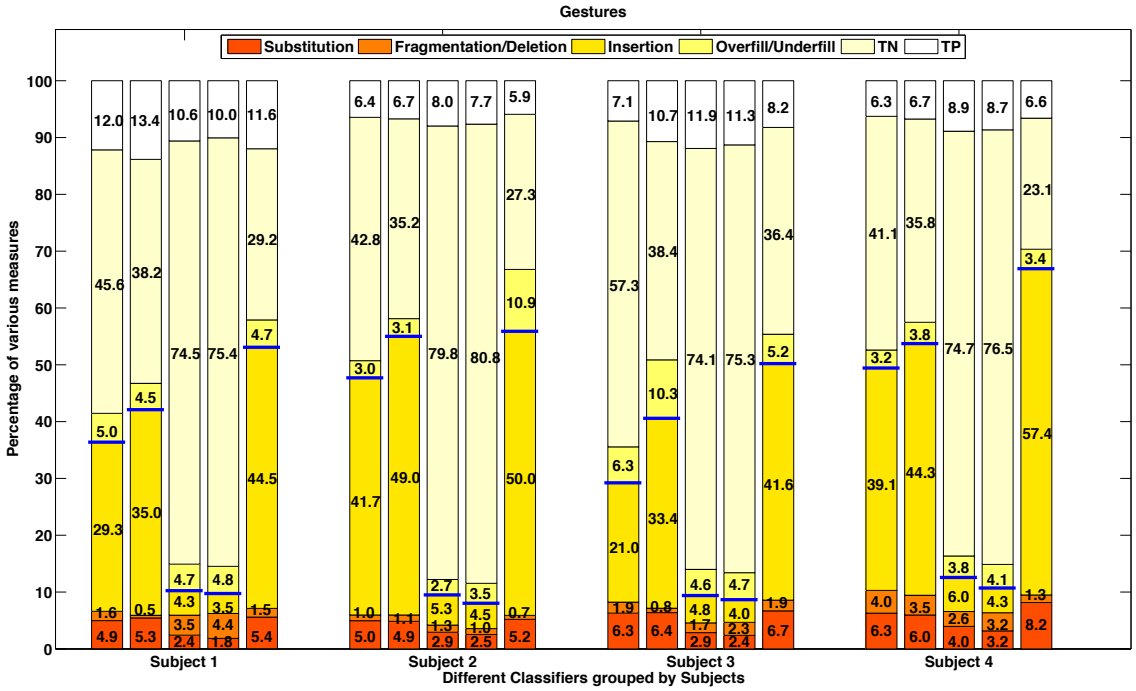


Figure 4: Gesture recognition - Use of a class-independent threshold to reject *Null* class samples (see section 7.3). Each group of five columns denotes the accuracy of LDA, QDA, 1-NN, 3-NN and NCC, respectively.

R., 2006. Using a live-in laboratory for ubiquitous computing research, in: Proc. Int. Conf. on Pervasive Computing, pp. 349–365.

van Kasteren, T., Noulas, A., Englebienne, G., Kröse, B., 2008. Accurate activity recognition in a home setting, in: Proceedings of the 10th international conference on Ubiquitous computing, ACM Press. pp. 1–9.

Kawaguchi, N., Ogawa, N., Iwasaki, Y., Kaji, K., Terada, T., Murao, K., Inoue, S., Kawahara, Y., Sumi, Y., Nishio, N., 2011. HASC challenge: gathering large scale human activity corpus for the real-world activity understandings, in: Proceedings of the 2nd Augmented Human International Conference, ACM, New York, NY, USA. pp. 27:1–27:5.

Pirkl, G., Stockinger, K., Kunze, K., Lukowicz, P., 2008. Adapting magnetic resonant coupling based relative positioning technology for wearable activitiy recogniton, in: Wearable Computers, 2008. ISWC 2008. 12th IEEE International Symposium on, pp. 47 –54.

Preece, S.J., Goulermas, J.Y., Kenney, L.P.J., Howard, D., Meijer, K., Crompton, R., 2009. Activity identification using body-mounted sensors – a review of classification techniques. Physiological Measurement 30, R1.

Roggen, D., Calatroni, A., Rossi, M., Holleczek, T., Förster, K., Tröster, G., Lukowicz, P., Bannach, D., Pirkl, G., Ferscha, A., Doppler, J., Holzmann, C., Kurz, M., Holl, G., Chavarriaga, R., Sagha, H., Bayati, H., Creatura, M., Millán, J.R., 2010. Collecting complex activity data sets in highly rich networked sensor environments, in: Seventh International Conference on Networked Sensing Systems, pp. 233–240.

Rohrbach, M., Amin, S., Andriluka, M., Schiele, B., 2012. A database for fine grained activity detection of cooking activities, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 1194–1201.

Saar-Tsechansky, M., Provost, F., 2007. Handling missing values when applying classification models. Journal of Machine Learning Research 8, 1623–1657.

Sagha, H., Digumarti, S., del Millán, J., Chavarriaga, R., Calatroni, A., Roggen, D., Tröster, G., 2011a. Benchmarking classification techniques using the opportunity human activity dataset, in: Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on, pp. 36 –40.

Sagha, H., Millán, J.d.R., Chavarriaga, R., 2011b. Detecting anomalies to improve classification performance in an opportunistic sensor network, in: 7th IEEE International Workshop on Sensor Networks and Systems for Pervasive Computing, PerSens 2011, Seattle. pp. 154–159.

Stiefmeier, T., Roggen, D., Ogris, G., Lukowicz, P., Tröster, G., 2008. Wearable activity tracking in car manufacturing. IEEE Pervasive Computing Magazine 7, 42–50.

Tenorth, M., Bandouch, J., Beetz, M., 2009. The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition, in: IEEE Int. Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS) at ICCV, pp. 1089–1096.

Tentori, M., Favela, J., 2008. Activity-aware computing for healthcare. IEEE Pervasive Computing Magazine 7, 51 –57.

Van Laerhoven, K., Gellersen, H., Malliaris, Y., 2006. Long-term activity monitoring with a wearable sensor node, in: BSN 2006. International Workshop on Wearable and Implantable Body Sensor Networks, pp. 171–174.

Ward, J.A., Lukowicz, P., Gellersen, H.W., 2011. Performance metrics for activity recognition. ACM Trans. Intell. Syst. Technol. 2, 6:1–6:23.

Webb, G., 1999. Decision tree grafting from the all tests but one partition, in: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. pp. 702–707.

Zappi, P., Stiefmeier, T., Farella, E., Roggen, D., Benini, L., Tröster., G., 2007. Activity recognition from on-body sensors by classifier fusion: Sensor scalability and robustness, in: 3rd International Conference on Intelligent Sensors, Sensor Networks, and Information Processing (ISSNIP), pp. 281–286.

Zinnen, A., Blanke, U., Schiele, B., 2009. An analysis of sensor-oriented vs. model-based activity recognition, in: International Symposium on Wearable Computers, IEEE Press. pp. 93–100.

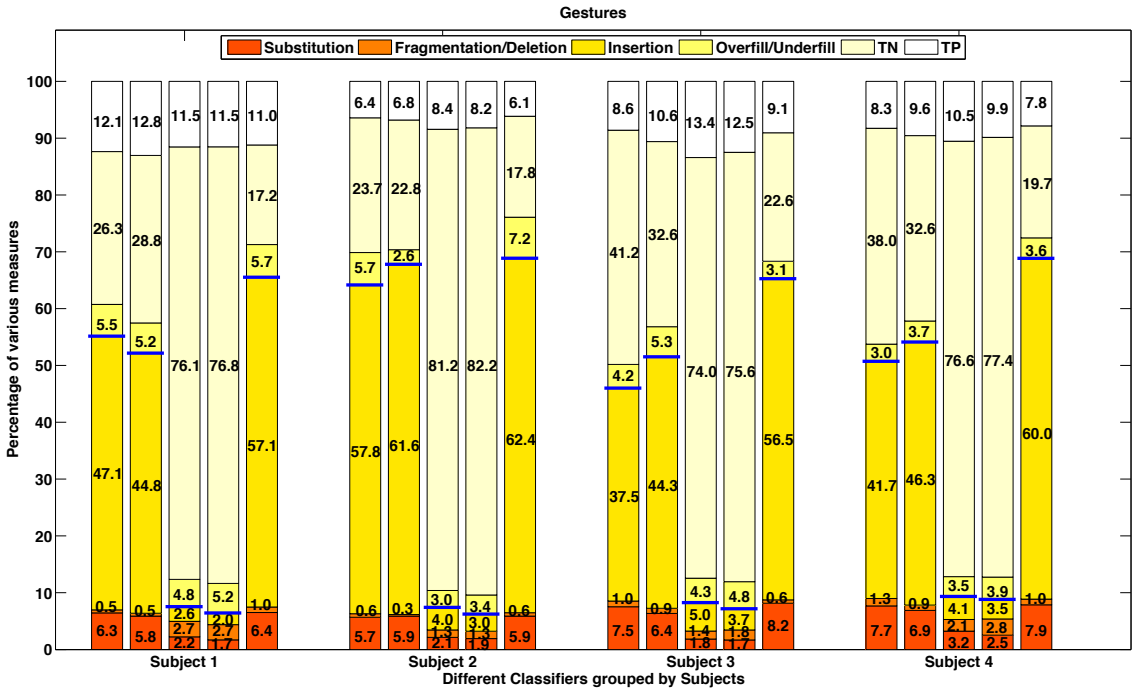


Figure 5: Gesture recognition - Using only motion-jacket sensors without rotational noise. Each group of five columns denotes the accuracy of LDA, QDA, 1-NN, 3-NN and NCC, respectively.

Video Still

Click here to download high resolution image