Benchmarking classification techniques using the Opportunity human activity dataset

Hesam Sagha, Sundara Tejaswi Digumarti,
José del R. Millán, Ricardo Chavarriaga
CNBI, Center for Neuroprosthetics,
Ecole Polytechnique Fédérale de Lausanne (EPFL)
hesam.sagha@epfl.ch

Alberto Calatroni, Daniel Roggen,
Gerhard Tröster
Wearable Computing Laboratory
ETH Zürich
Switzerland

***Abstract*—Human activity recognition is a thriving research field. There are lots of studies in different sub-areas of activity recognition proposing different methods. However, unlike other applications, there is lack of established benchmarking problems for activity recognition. Typically, each research group tests and reports the performance of their algorithms on their own datasets using experimental setups specially conceived for that specific purpose. In this work, we introduce a versatile human activity dataset conceived to fill that void. We illustrate its use by presenting comparative results of different classification techniques, and discuss about several metrics that can be used to assess their performance. Being an initial benchmarking, we expect that the possibility to replicate and outperform the presented results will contribute to further advances in state-of-the-art methods.**

***Index Terms*—Human activity recognition, Benchmark , Body sensors networks, Opportunity dataset,.**

I. Introduction

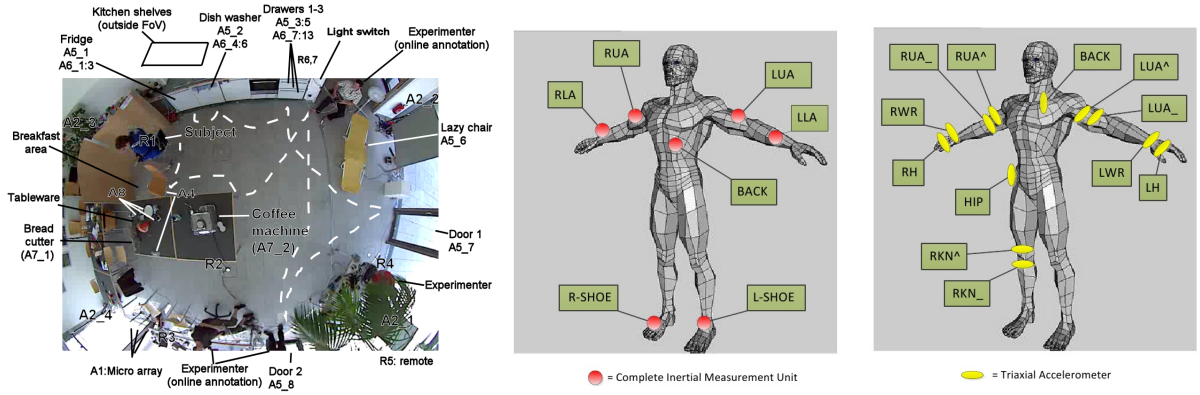
Recently, there has been an increasing attention towards human activity recognition –using on-body, object-placed or ambient sensors– fostered by applications in health care [1], [2], assistive technologies [3], manufacturing [4], or gaming (e.g. Microsoft Kinect or [5]). These applications apply machine learning techniques to classify signals gathered by different types of sensors. Indeed, this field typically requires to deal with high-dimensional, multimodal streams of data that are characterized by a large variability (e.g. due to changes in the user's behavior or as a result of noise). Therefore, several challenges arise at the different processing stages from the feature selection and classification [6], [7], to sensor and decision fusion [8], as well as fault-tolerance [9], [10], [11]. Moreover, real-life deployments are required to detect when no relevant action is performed (i.e. null class). Therefore, there is a need for methods able to spot the specific time points when relevant actions are being executed [4], [12].

However, unlike other applications, there is lack of established benchmarking problems for activity recognition. Typically, each research group tests and reports the performance of their algorithms on their own datasets using experimental setups specially conceived for that purpose. For this reason, it is difficult to compare the performance of different methods or to assess how a particular technique will perform if the experimental conditions change (e.g. in case of sensor failure or changes in sensor location). We argue that there is a need for common databases that allow the comparison of different machine learning algorithms on the very same conditions. Such database would enable the replication of the testing procedures for different approaches, and should capture the variability that characterizes real-world activity recognition tasks. Moreover, it should be flexible enough to emulate different experimental setups and recording modalities [13].

The Opportunity dataset is intended to address these issues by providing a large recording of realistic daily life activities in a sensor rich environment [14], [15]. Moreover a subset of this dataset is the basis of the activity recognition challenge (http://www.opportunity-project.eu/challenge) aimed at comparing different systems –developed by several research groups– addressing the recognition of gestures and modes of locomotion using body-worn sensors. This paper illustrates the use of the dataset for comparing different techniques by presenting a benchmarking study of four well-known classification techniques, namely k-NN, NCC as well as Gaussian classifiers (LDA and QDA). Moreover, in order to assess the robustness of these methods, we also report classification performance on data where rotational noise has been added.

II. Opportunity Dataset

The Opportunity dataset was acquired from 12 subjects while they are performing morning activities and includes 72 sensors of 10 modalities in 15 wireless and wired networked sensor systems in the environment, objects and the body, as shown in Fig 1(a) [14]. For each subject there are five daily activity sessions and one drill session which has about 20 repetitions of some pre-defined actions. Data was manually labeled for modes of locomotion, gestures and high-level activities by at least two different persons [15]. In this paper, we use a subset of the recording corresponding to 4 subjects and focus on recognition of gesture and modes of locomotion using body-worn sensors (class labels are presented in Table I), while sensor locations are shown in Fig. 1(b),1(c). Moreover, in order to test robustness of the methods, rotational noise was added to the recordings of subject 4. Several applications may suffer this type of noise. For instance, in some cases when the user should re-attach the sensors over different days it is unrealistic to expect him/her to place them always in the same orientation. Similarly, sensors on mobile phones placed in a



(a) (b) (c)

Fig. 1. (a) Recording environment of the Opportunity dataset. (b) Location of the on-body IMU sensors. (c) Location of the bluetooth accelerometers.

pocket may easily rotate as the person carries it over the day. The rotation angle is at maximum 60 degrees in any direction and all the sensors in the IMU (accelerometer, gyro, and magnetic sensors) are affected. The rotation process is started at a random time for each IMU. Following the guidelines of the proposed challenge, the wired sensors located in the upper body were used for the classification of this recording.

III. Classification Methods

We present comparative results for activity recognition using simple and standard classification methods. It should be noticed that the goal of this study is to provide a baseline characterization of the difficulty of the task, rather than aiming to achieve the highest possible performance. Indeed, these results should be used to assess how much improvement can be obtained by using different, more complex approaches for feature selection and classification.

In order to deal with the missing data in the wireless sensors, although more complex methods are available (c.f. [16]), we opt for the simple repetition of the last available value. Taking into account that the data is not segmented, we perform classification in a sliding window of 500ms, with steps of 250ms. Experiments were performed using the mean value of the sensor readings as feature, as well as both the mean and the variance. Note that, the average length of gestures is about 3.5 seconds, and the shortest gesture found in the dataset lasts about 0.5 second, so the choice of half a second is a good deal to have a high speed of recognition without missing classes. Moreover, for some classifiers (NCC, LDA, QDA) a rejection mechanism was implemented to identify samples that correspond to the Null class. We tested the following classifiers,

K-nearest neighbors (k-NN). We perform simulations using k=1 and k=3. Since all the feature points are stored and the Null class is explicitly modeled there is no need of rejection procedure.

Nearest Centroid Classifier (NCC). Since the Null-class samples are scattered in the feature space, it is not reasonable to treat them as another class, and we implemented instead a threshold-based rejection procedure. We initially train the classifier using only the activity labels (i.e. excluding Null-class samples of the training set). Then, using the whole dataset we estimate class-specific thresholds on the distance to the class center that maximize the accuracy (F-measure).

Linear Discriminant Analysis (LDA). This is a Gaussian statistical method that assumes that class features have a normal distribution and all classes have the same covariance matrix. Similar to NCC, we implemented a rejection method, where the thresholds are defined on the posterior probabilities.

Quadratic Discriminant Analysis (QDA). This classifier also assumes that classes are normally distributed but does not assume identical class covariances. Therefore, it results in a quadratic discriminative function, instead of a linear function. The same rejection method as LDA is used.

TABLE I
Class labels for both models of locomotion and gestures recognition.

Modes of locomotion				
Null	Stand	Sit	Walk	Lie
Gestures				
Null	clean Table	open Drawer1	close Drawer1	
open Dishwasher	close Dishwasher	open Drawer2	close Drawer2	
open Fridge	close Fridge	open Drawer3	close Drawer3	
open Door1	close Door1	open Door2	close Door2	
move Cup				

IV. Performance Measures

There are several ways to assess the performance of an activity recognition system. However, the choice of an appropriate measure is not trivial as these measures may reflect some specific qualities of the system while hiding or misrepresenting others (c.f. [17]). This becomes even more important when dealing with real-life data where 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 labeling). Similarly, during periods labeled as null –denoting when none of the class actions is performed– it cannot be assumed that the person remained still; indeed, most of the time s/he is performing another action or in a transition

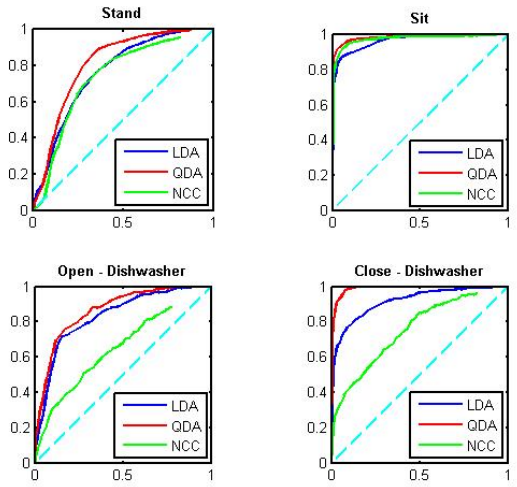


Fig. 2. ROC curves of NCC, LDA and QDA classifiers. Four gesture classes are shown: Stand, Sit, Open and Close Dishwasher.

from one action to another. Furthermore, continuous recordings may be highly unbalanced with one of the classes being overrepresented with respect to the others. This is the current case for gesture recognition where 'Null' class represents about 80% of the recorded data (76%, 82%, 76% and 78% for subjects 1 to 4, respectively).

The simplest performance measure is the accuracy (acc = correct predicted/number of samples), which is highly affected by the sample distribution across activity classes. Alternatively, the F-measure –taking into account the precision and recall for each class– can give a better assessment of performance. 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 is the proportion of samples of class i ($w_i = n_i/N$). Similarly, the area under the curve (AUC) in the ROC space can also be used as a performance measure (c.f. Figure 2). As for the F-measure, the class imbalance can be taking into account by weighting the AUC for each class by its prevalence on the data [18].

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

In addition, as actions onset and offset times are not precisely defined, misalignment of output labels (e.g. early detection of an action onset) may be wrongly considered as classification errors. Ward et al. propose to explicitly quantify the system performance taking all these aspects into account [19]. They characterized different types of errors as follows (listed in increasing order of importance),

1) Overfill: when the start and stop time of predicted labels are less and greater than actual time, respectively.
2) Underfill: when the start and stop time of predicted labels are greater and less than actual time, respectively.
3) Merge: recognizing Null label as a label of an action in the middle of its occurrence.
4) Insertion: when an activity is recognized while there is no activity.
5) Fragmentation: predicting Null in between of an action.
6) Deletion: the predicted label is recognized as Null class, but in fact it is an activity going on.
7) Substitution: The predicted and actual labels are not Null but they are not the same.

Note that overfill and underfill may not necessarily correspond to recognition errors but the result of alignment variation with the label used as ground truth.

V. Results

We report the recognition performance using the classification methods presented in Section III. Please note that results on subjects 2 and 3 correspond to the tasks A, and B2 of the activity recognition challenge (modes of locomotion and gesture recognition, respectively), and results on subject 4 corresponds to the Task C (noisy data). Surprisingly, using both the mean and variance of the signal does not improve performance as compared to use only the mean values. In the following, we report only the results obtained using the mean as feature. Table II shows the weighted F-measure as well as accuracy. We present two ways of computing the F- measure, either including or not the Null class[1]. Overall, the best performance was achieved by the kNN classifier for recognizing both locomotion and gestures, followed by the Gaussian classifiers. From the table, it can also be seen the effect of the class imbalance, as the inclusion of the Null class leads to an overestimation of the accuracy in the gesture recognition problem. The same effect is observed when comparing the weighted AUC for NCC, LDA and QDA, as shown in Table III.

The detailed measures proposed by Ward et al. are shown in Figure 3. This confirms the results obtained with the F-measures that point out the higher performance of the kNN classifiers, even with noisy data. When recognizing modes of locomotion, these classifiers had a smaller rate of overfill and underfill than other classifiers, suggesting it accurately captures the on/offset of the actions. Unsurprisingly, this percentage increases for subject 4 that has noisy data and only sensors on the upper torso are available. Regarding gesture recognition, the advantage of kNN results from its reduced level insertions errors, suggesting that the threshold-based rejection mechanism is not always able to discriminate the Null class. This is probably due to the overlapping in the feature distribution.

VI. Conclusion

This paper presents a comparative study of classification techniques for activity recognition. We assessed performance using different measures for standard classification techniques

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

TABLE II
ACCURACY AND WEIGHTED F-MEASURE.

Modes of Locomotion

	Accuracy					F-measure (Incl Null class)					F-measure (Without Null class)				
Classifier	S1	S2(A)	S3(A)	S4	Avg	S1	S2(A)	S3(A)	S4	Avg	S1	S2(A)	S3(A)	S4	Avg
LDA	0.66	0.64	0.68	0.44	0.60	0.64	0.64	0.68	0.43	0.60	0.75	0.70	0.74	0.53	0.68
QDA	0.71	0.67	0.72	0.47	0.64	0.67	0.63	0.71	0.45	0.62	0.80	0.74	0.79	0.59	0.73
1-NN	0.84	0.85	0.84	0.77	0.82	0.84	0.85	0.83	0.77	0.82	0.85	0.85	0.85	0.76	0.83
3-NN	0.85	0.86	0.85	0.78	0.83	0.85	0.86	0.84	0.77	0.83	0.86	0.86	0.86	0.77	0.84
NCC	0.62	0.59	0.55	0.41	0.54	0.60	0.58	0.56	0.41	0.54	0.69	0.67	0.62	0.50	0.62

Gesture recognition

	Accuracy					F-measure (Incl Null class)					F-measure (Without Null class)				
Classifier	S1	S2(B2)	S3(B2)	S4(C)	Avg	S1	S2(B2)	S3(B2)	S4(C)	Avg	S1	S2(B2)	S3(B2)	S4(C)	Avg
LDA	0.58	0.44	0.64	0.54	0.53	0.64	0.54	0.69	0.60	0.62	0.34	0.26	0.33	0.19	0.28
QDA	0.52	0.35	0.62	0.48	0.49	0.57	0.44	0.68	0.55	0.56	0.32	0.25	0.39	0.19	0.29
1-NN	0.82	0.84	0.85	0.81	0.83	0.82	0.84	0.85	0.81	0.83	0.53	0.47	0.62	0.47	0.52
3-NN	0.83	0.85	0.85	0.83	0.84	0.82	0.85	0.85	0.82	0.83	0.52	0.49	0.62	0.48	0.53
NCC	0.42	0.39	0.49	0.27	0.39	0.48	0.48	0.55	0.33	0.46	0.30	0.21	0.29	0.15	0.24

TABLE III
WEIGHTED AREA UNDER THE CURVE (AUC)

Modes of Locomotion

Classifier	S1	S2(A)	S3(A)	S4	Avg
LDA	0.77	0.71	0.75	0.63	0.72
QDA	0.82	0.79	0.83	0.68	0.78
NCC	0.72	0.68	0.71	0.59	0.68

Gesture recognition

Classifier	S1	S2(A)	S3(A)	S4	Avg
LDA	0.89	0.81	0.87	0.89	0.86
QDA	0.92	0.88	0.90	0.91	0.90
NCC	0.79	0.76	0.82	0.78	0.79

such as k-NN, NCC, LDA, and QDA. A particularly important issue that we observed was the effect of class imbalance in the evaluation of F-measures and overall performance. Indeed, as we deal with continuous, unsegmented data, the Null class is overrepresented in the dataset. Furthermore, these samples may contain activity that overlaps with some of the selected classes. This aspect has to be taken into account when designing a system (e.g. by including risk function in the optimization of the classifier parameters). Alternatively, dedicated methods to automatically segment can be developed (as proposed by Task B1 in the activity recognition challenge).

Based on the Opportunity activity recognition challenge, this work 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, and one of the subjects has a different sensor configuration and noisy data. The selection of the compared methods aims at providing a baseline performance. Since the data is publicly available, these baseline results can be later used by other researchers to assess how much improvement is achieved when more complex techniques are applied. We expect that the possibility to replicate and outperform these results will contribute to further advances in state-of-the-art methods.

VII. ACKNOWLEDGEMENTS

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

[1] E. Tapia, S. Intille, and K. Larson, "Activity recognition in the home using simple and ubiquitous sensors," in *Pervasive Computing*, ser. Lecture Notes in Computer Science, A. Ferscha and F. Mattern, Eds. Springer Berlin / Heidelberg, 2004, vol. 3001, pp. 158–175.

[2] M. Tentori and J. Favela, "Activity-aware computing for healthcare," *Pervasive Computing, IEEE*, vol. 7, pp. 51 –57, 2008.

[3] M. Bächlin, M. Plotnik, D. Roggen, N. Giladi, J. M. Hausdorff, and G. Tröster, "A wearable system to assist walking of parkinson's disease patients." *Methods Inf Med*, vol. 49, pp. 88–95, 2010.

[4] T. Stiefmeier, D. Roggen, G. Ogris, P. Lukowicz, and G. Tröster, "Wearable activity tracking in car manufacturing," *IEEE Pervasive Computing Magazine*, vol. 7, pp. 42–50, 2008.

[5] H. Kang, C. W. Lee, and K. Jung, "Recognition-based gesture spotting in video games," *Pattern Recognition Letters*, vol. 25, pp. 1701–1714, 2004.

[6] S. J. Preece, J. Y. Goulermas, L. P. J. Kenney, D. Howard, K. Meijer, and R. Crompton, "Activity identification using body-mounted sensors – a review of classification techniques," *Physiological Measurement*, vol. 30, no. 4, p. R1, 2009.

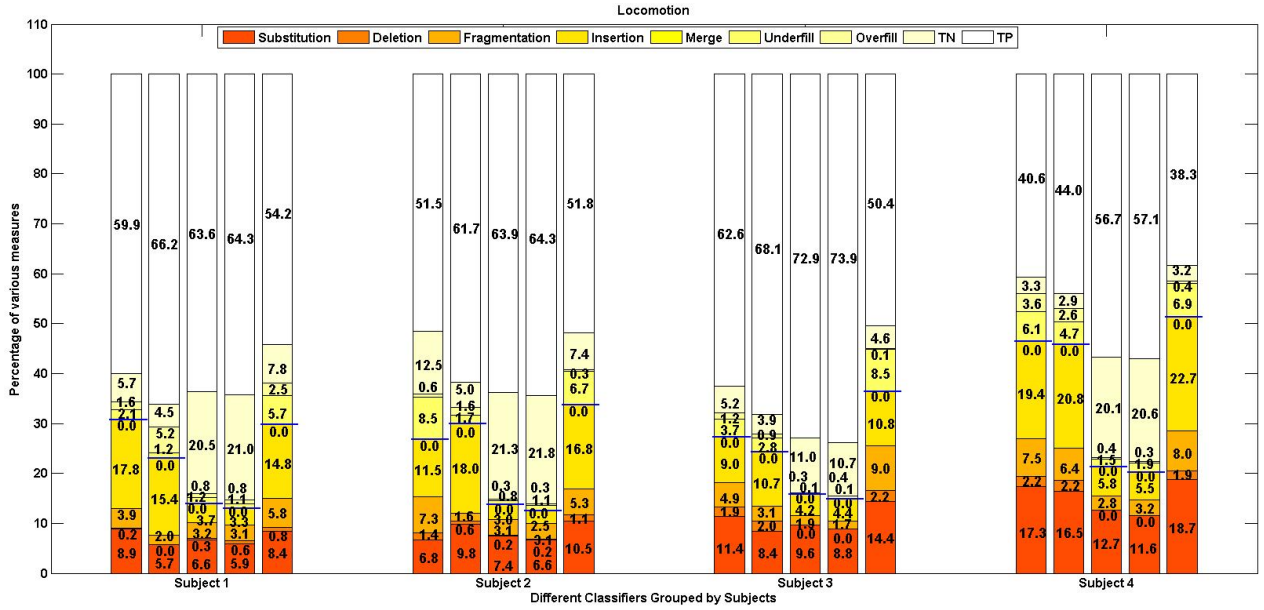
[7] D. Figo, P. C. Diniz, D. R. Ferreira, and J. a. M. Cardoso, "Preprocessing techniques for context recognition from accelerometer data," *Personal Ubiquitous Computing*, vol. 14, pp. 645–662, 2010.

[8] P. Zappi, T. Stiefmeier, E. Farella, D. Roggen, L. Benini, and G. Tröster., "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)*, 2007, pp. 281–286.

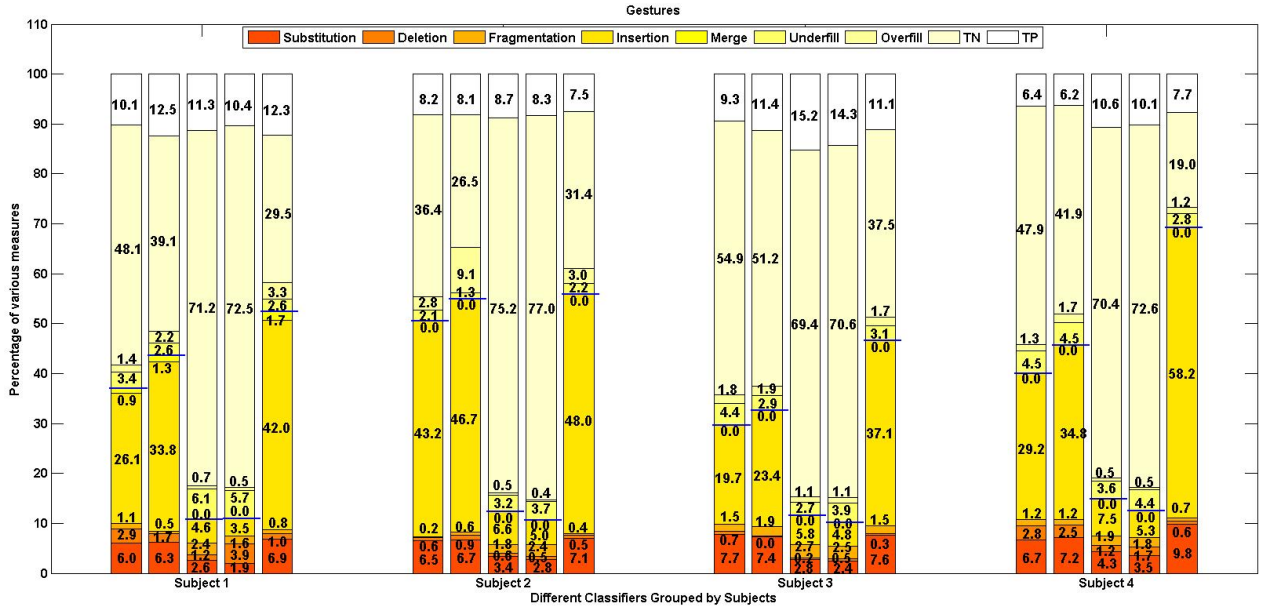
[9] K. Kunze and P. Lukowicz, "Dealing with sensor displacement in motion-based onbody activity recognition systems," in *Proceedings of the 10th international conference on Ubiquitous computing*, 2008, pp. 20–29.

[10] H. Bayati, J. d. R. Millán, and R. Chavarriaga, "Unsupervised adaptation to on-body sensor displacement in acceleration-based activity recognition," in *IEEE International Symposium on Wearable Computers, ISWC*, 2011.

[11] H. Sagha, J. d. R. Millán, and R. Chavarriaga, "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, March 2011.



(a)



(b)

Fig. 3. Recognition performance using the Ward's measures [19]. (a) Modes of locomotion and (b) Gesture recognition. 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, leading to a performance decrease.

[12] H. Junker, P. Lukowicz, and G. Tröster, "Continuous recognition of arm activities with body-worn inertial sensor," in *Eighth International IEEE Symposium on Wearable Computers*, 2004, pp. 188–189.

[13] R. Chavarriaga, H. Sagha, H. Bayati, J. d. R. Millán, D. Roggen, K. Förster, A. Calatroni, G. Tröster, P. Lukowicz, D. Bannach, M. Kurz, G. Hölzl, and A. Ferscha, "Robust activity recognition for assistive technologies: Benchmarking machine learning techniques," in *Workshop on Machine Learning for Assistive Technologies -at (NIPS)*, 2010.

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

[15] P. Lukowicz, G. Pirkl, D. Bannach, F. Wagner, A. Calatroni, K. Förster, T. Holleczek, M. Rossi, D. Roggen, G. Tröster, J. Doppler, C. Holzmann, A. Riener, A. Ferscha, and R. Chavarriaga, "Recording a complex, multi modal activity data set for context recogntion," in *Workshop on Context-Systems Design, Evaluation and Optimisation at ARCS*, 2010.

[16] M. Saar-Tsechansky and F. Provost, "Handling missing values when applying classification models," *Journal of Machine Learning Research*, vol. 8, pp. 1623–1657, 2007.

[17] H. Junker, J. Ward, P. Lukowicz, and G. Tröster, Eds., *Benchmarks and a Data Base for Context Recognition.*, 2004, ISBN 3-9522686-2-3.

[18] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861 – 874, 2006.

[19] J. Ward, P. Lukowicz, and G. Tröster, "Evaluating performance in continuous context recognition using event-driven error characterisation," in *Location- and Context-Awareness*, M. Hazas, J. Krumm, and T. Strang, Eds. Springer Berlin / Heidelberg, 2006, vol. 3987, pp. 239–255.