User Independent, Multi-Modal Spotting of Subtle Arm Actions with Minimal Training Data

Gerald Bauer*, Ulf Blanke†, Paul Lukowicz* and Bernt Schiele‡
*Embedded Intelligence, German Research Center for Artificial Intelligence, Kaiserslautern, Germany
Email: {gerald.bauer, paul.lukowicz}@dfki.de
†Wearable Computing Laboratory, ETH, Zurich, Switzerland
Email: ulf.blanke@ife.ee.ethz.ch
‡Computer Vision and Multimodal Computing, Max Planck Institute for Informatics, Saarbrücken, Germany
Email: schiele@mpi-inf.mpg.de

***Abstract*—We address a specific, particularly difficult class of activity recognition problems defined by (1) subtle, and hardly discriminative hand motions such as a short press or pull, (2) large, ill defined NULL class (any other hand motion a person may express during normal life), and (3) difficulty of collecting sufficient training data, that generalizes well from one to multiple users. In essence we intend to spot activities such as opening a cupboard, pressing a button, or taking an object from a shelve in a large data stream that contains typical every day activity. We focus on body-worn sensors without instrumenting objects, we exploit available infrastructure information, and we perform a one-to-many-users training scheme for minimal training effort. We demonstrate that a state of the art motion sensors based approach performs poorly under such conditions (Equal Error Rate of 18% in our experiments). We present and evaluate a new multi modal system based on a combination of indoor location with a wrist mounted proximity sensor, camera and inertial sensor that raises the EER to 79%.**

***Keywords*-Activity Spotting; ADL; Wearable Sensors; Hand Mounted Camera; Multi-Modal Sensing**

I. Introduction

Activity spotting aims to detect specific individual actions in a continuous stream of arbitrary activity. Often the actions of interest constitute only a small part of the overall signal, and are embedded in a "everything else" NULL class for which building reliable models is impractical. A particularly difficult version of the spotting problem relates to actions that are determined by simple and short hand or arm gestures such as pressing a button, turning a knob, picking something up or putting it away. The NULL class then consists of "all the other arm motions that a person may have" – including motions that are very similar to the relevant actions.

Much previous work has been investigating the use of body mounted motion sensors (accelerometers, gyroscopes, magnetic field) for that purpose. Two factors have been shown to be critical for the success of such approaches:

1) The presence of distinct and characteristic motion segments that are unlikely to occur in the NULL class. In particular very simple actions such as pressing a button or pulling a lever often lack such characteristic segments. In the following we illustrate this by showing that a state of the art motion based recognition system performs very poorly on such a data set.
2) Availability of sufficient training data, preferably on a user specific basis. While easy in lab experiments this is a significant hurdle for the practical deployment of activity recognition systems. Real life users expect their systems to work out of the box that do not require tens of repetitions to be provided for training.

In this paper we investigate how extending the body-worn sensor system beyond motion sensors can contribute to overcome the restrictions mentioned above. We envision a system that does not rely on on large amounts of statistically significant training data from the user. Instead the models are constructed from "one time" measurements performed by the person installing the system.

II. Related Work

The recognition of daily life activities and interactions with objects is a well known research topic. Some approaches use eye tracking systems to detect object interactions [15] [10]. In [15] time frames are marked as interesting, if the person stares at a object for a longer time duration. Based on a huge set of training images for which each object was covered from all possible views, SIFT-matching is used to identify the object. Other systems use wearable cameras and microphones to recognize a person's situation [15] [10]. However they consider not specific activities but the coarse location of a person. There are also many approaches based on radio systems (e.g. [11] [6]). The big disadvantage here is that the reading range is often limited and each object has to be instrumented. Other approaches are based on the assumption that every object is able to provide its state with binary switches [14] or infrared sensors [12]. In [13] a similar approach to our idea is presented. This system uses also a wrist-worn camera in combination with other body-worn sensor systems. However it uses characteristic features based on color histograms, whereas our system uses the shape of an object and hence it is more robust to changing

978-1-4673-5077-8/13/$31.00 ©2013 IEEE

light conditions. Moreover, it requires a significant amount of training data.

III. Experimental Setup

We selected 16 different object interactions evolved from 30 activities (see Table I). The activities have been recorded in a continuous data stream within a real office environment (see Figure 1) including four rooms (student room, printer room, kitchen, office) and a corridor. We hired 6 students to repeat the activities in a randomly generated sequence all in all 27 times (713 performed object interactions, almost 7 hours of recording) during the normal working time. In this way we can guarantee a real multi user environment, where also non-participants have been present and have also performed activities such as using the coffee machine or the printer.

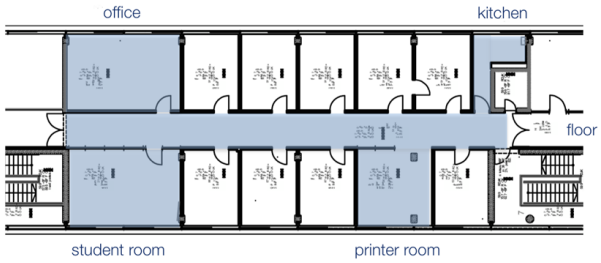

Figure 1. Floor plan: The blue area markes the monitored region.

Besides the defined activities our participants performed the following background activities: Clean whiteboard (student room), write something on whiteboard (student room), open/close door (student room), count coins on table (office), rifle items on a table (office), open / close window (office), point with finger at a wall mounted map (office), admire picture on wall (office), drink from glass (kitchen), take milk from fridge (kitchen), stir up coffee in cup (kitchen), sit down and read newspaper (kitchen), clean table (kitchen), put printout on wall (printer room). Additionally participants had to behave as they do normally while performing the predefined activities. In this way we recorded almost 7 hours of data of which 91% belongs to the background class.

IV. Sate-Of-The-Art: Using Inertial Sensor to Detect Object Interactions

We employ a baseline activity recognition system based on common steps of segmenting continuous data into regions, calculating features per segment (or region), and feeding the features together with labels into the learning procedure of a classification model. For the classification task, the continuous data stream is segmented in potential candidates for an activity. Then, features are extracted and scores are obtained for the trained classes, respectively the activities.

Table I
Performed Object Interactions Grouped By Rooms: Kitchen, Printer Room, Student Room and Office

Object	Activities (Repetitions)
Microwave	Open (27), Close (27), Start (11), Clean (16)
Coffee Machine	Make Espresso (14), Make Coffee (13)
Power Socket	Connect Cable (27)
Cupboard	Open (27), Close (27)
Wall Cupboard	Open (27), Close (26)
Ethernet Connector	Connect Cable (30)
Water Tap	Fill Big Cup (14), Fill Small Cup (13)
Battery Charger	Put Battery (15), Remove Battery (14)
Laser Printer	Take Printout (12), Push Button (15)
Color Printer	Take Printout (15), Push Button (12)
Climatic Control	Change State (27)
PC	Turn On (55)
Scanner	On (29), Off (27), Scan Document (27)
Air Conditioner	On (27), Off (27)
Light-Shutter Switch	Use Light / Shutter Button (28 / 28)
Ring Binder	Take From Shelf (28), Put Back (28)

Basic Segmentation Procedure: Several segmentation techniques exist (e.g. [1, 18]) to partition a continuous data stream into candidates for an activity. The most common approach though is to use a sliding window approach with fixed window length, which is also used in this work. We estimate the window length from mean μ of the activity duration distribution. While a fixed sliding window size might not be the optimal choice and the choice of length can influence the recognition [7] we choose this method as a baseline which proved to work in numerous works targeting a variety of activities [2, 5, 8, 16].

Feature Calculation: Given a list of segments from above we calculate for each segment common features such as mean and variance, as well as frequency space based features. We standardize the feature space of the training set. For the test data we use the standardization parameters from the training set.

Classifier: For multi-class classification we use support vector machines with a radial basis function as kernel. As multi-class SVM we use a one-vs-one learning scheme. We experimentally obtain regularization parameters from the training set, which consist of 15 repetitions per activity performed by a person that has not participated in the data recording. In this way our trained system is person independent. During learning, we extract random segments from the background class, which we add to the training data. To this end we extract an equal number of negative samples compared to positive samples of all classes together. The model is then trained for probability estimates between 0 and 1 for each class. The classification step returns a normalized score vector per segment, where each element contains the score for a respective activity. Since we have

multiple overlapping windows from the segmentation step we perform an additional non-maximum suppression. To this end, for each timeframe all overlapping windows are selected. For each activity the maximum score is determined and kept as final score for the activity for the timeframe. As a result we obtain for each timeframe the label of the activity that achieved the highest score.

Final Segmentation Procedure: Based on the scores per timeframe we calculated new segments for each class. Therefore we defined segments with a fixed minimum (more than about 1/3 second) and maximum length (less than 20 seconds) in which all scores are above a certain SVM score threshold. We also fused segments that are close to each other (we evaluated several thresholds between 0 seconds and 1 second, in 1/5 second steps).

Baseline Results: As can be seen from Table II the system can achieve reasonable recall (80%) however because of the lack of really characteristic motion segments it has a precision of only 3% (EER of 18%). In addition it requires a considerable amount of training data (contrary to our system as described below).

V. Multi-Modal Sensor Approach

A. On-Body Approach: The Basic System

Our basic system consists of three different body worn, affordable and mainstream sensor systems: A forearm worn camera (Logitech C910, 640x480pixels, about 17 frames a second) in combination with a infra-red distance sensor (www.toradex.com, sampling rate about 10 Hz, range: 10 cm to 1 m, infra-red beam opening of 2 degree) and three Xsens inertial sensors placed on forearm, upper arm and on the back (see Figure 2).

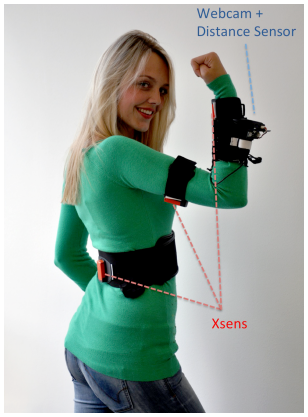

Figure 2. Body-worn sensors: Lower arm mounted camera and distance sensor, Xsens acceleration sensors on forearm, upper arm and back.

To spot relevant intervals and to identify a specific object interaction we perform the following three steps.

1) Spotting interesting time sequences using the ifra-red distance sensor: Hand actions in general involve object manipulations, which means that object interactions can only occur when the hand is close to the object. So we spot all time sequences TS_i where the hand of the person is "close" to an object according to the infrared proximity sensor. To find interesting time sequences we chose a distance threshold which is about 10 cm away from the tip of the persons finger. Note that this step involves no statistical training and only a single measurement of the proximity sensor placement on the arm is needed.

2) Assigning relevant objects to a time sequence: We use inertial sensors that provide a global orientation in Euler angles. We calculate a simple body model that allows us to determine the height of the users right forearm. To this end the system must be configured using the length of the users forearm, upper arm, body as well as legs. The hand height information is used to split each TS_i in several sub sequences TSS_{i_j}, where the maximum hand deviation is less than 10 cm. After that we assign a set of relevant objects o to each TSS_{i_j}. A TSS_{i_j} contains an object o if the maximum deviation between the average hand height of TSS_{i_j} and the pre-configured height of the object o is between +/- 30 cm (In this way we cover both: the inaccuracy of the hand height calculation as well as hand height variation while interacting with an object). This measurement is done in a single step without involving statistical training.

3) Image based object recognition: So far we got several TSS, each containing a list of possible objects. We next use a computer vision object recognition algorithm based on SVM and HOG features (sliding window, block size 8x8) to identify the object the user is currently interacting with. Therefore all images within a TSS (We added +/- 1 second as the camera is not able to capture the whole object while the person is close to it) are analyzed using SVMs for each object class and the class with the biggest average SVM score is selected. To reduce false classifications we define a threshold, which must be exceeded if the object should be taken into account. To train a SVM we used only one single image per object from which 80 training images have been artificially generated by adjusting the brightness using a gamma filter. Thus, again the training amounts to a single measurement (taking one photo). In addition we assume that a person walks once through the office space recording random images as examples of the NULL class. During the evaluation process each image is rotated from -90 to 90 degrees and scaled between -0.5 and +4.0 in steps of 0.05.

B. Additional Sources of Information

1) Forearm Location (AH): To get the location of the lower right arm we measure the 3D magnetic field provided by the Xsens sensors. As reference data, we record the magnetic field around each object for only some seconds

(again, not statistical training). We compare the current magnetic field within a TSS with the one recorded for a specific object and using a distance metric on the magnetic field vector in combination with a time duration feature we reduce the amount of possible objects for each TSS.

2) Time Related Features (TF): We remove all TSS if the duration is much longer (finally 8 seconds) as the standard object interaction duration.

3) Modes of Locomotion (MoL): Using the acceleration sensor from a smartphone that is carried in the users pocket we are recognizing walking-standing activities using a standard technique based on the variance of the 3D acceleration. During a walking time period we assume that the person is not performing any activity and we don't take into account the belonging set of images.

4) No Hand Movement (NHM): We assume that just before the users perform an activity (like pushing the light button) the hand is moving quite fast to touch the object. So we reject all images where the users hand shows almost no movement.

C. Optional Infrastructure Sensors

1) Room Level Location (RLL): In this paper we choose a standard BT based approach and so one Bluetooth beacon was mounted in each room at a random place. We use a smartphone (carried in the pocket of the user) to scan for reachable Bluetooth beacons. Finally we assign a specific room to each TSS. In this way we reduce the list of possible classes by removing all objects that are not located in the current room (Note: light-shutter actuators appear in each room).

2) Regions of Interest (ROI): Each room has been equipped with a ceiling mounted fish-eye camera to monitor the activity within pre-defined so called regions of interest - ROI (see Figure 3). Each ROI can include a single object or even a group of objects (when objects are located close to each other). All in all we defined 11 ROI containing the 16 different objects. To determine if somebody is inside a ROI we use a standard approach based on difference images of two consecutive grey-scale camera images. If a person is within a ROI during a TSS, only objects that belong to this ROI are considered. The fact that several people are moving or acting within pre-defined ROIs and we assume that this activity was performed by our test person may lead to many false detections. In [3] a system is described to overcome this problem and hence our system can be even improved.

3) Operating Mode of Objects (OOM): To recognize the operating mode of a device we used a system similar to the one described in [4] [9]. For each TSS we remove the electronic device / the water tap if the operating mode of the device has not changed within the TSS. Due to the recognition delay of the referenced systems we tolerated a time difference of 3 seconds for electronic devices and of 2 seconds for the water tap. The following devices have been taken into account: microwave, water tap, pc, scanner, air conditioner, battery charger and coffee machine.

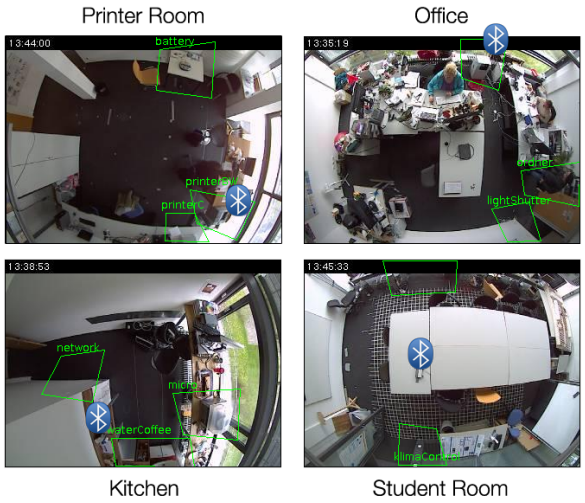

Figure 3. Room monitoring using ceiling mounted fish-eye cameras. The pre-defined ROIs (green rectangles) as well as the position of the Bluetooth beacons are shown.

VI. Evaluation and Discussion

Table II summarizes the recognition results for different combinations of sensors and methods described above (EER was evaluated by adjusting the SVM threshold). All system parameters have been optimized to reach the highest recall. We evaluate the quality of all systems as a function of a threshold on calculated SVM scores, which is used to reject false classifications. As image processing is highly time and power consuming, we investigated the amount of required classification steps and the number of analyzed images, aside from the quality of the recognition. To calculate recall and precision we count every overlap between a recognition output and a corresponding label as a true positive and every recognition output without a corresponding label as false positive. Furthermore, we use a more detailed evaluation method based on events and time samples (see [17], Figure 4). When using inertial sensors only it is almost impossible to reach a sufficient recognition accuracy (EER of 18%). On the other hand even our basic system based on a hand worn camera is able to provide a reasonable starting point for further sensor fusion (recall of 67%, precision of 26% and a EER of 43%). When analyzing the influence of the hand height tolerance thresholds (introduced in V-A2) we found out that a value pair of -25/10 cm provides the best result. In this way a recall of 75% and a precision of 22% was achieved (EER of 47%). Thereby 1.566.872 classification steps have been performed and 205.132 images have been analyzed.

When combining our basic vision setup with additional sensor systems we can see that location (ROI, RLL and

Table II
EVALUATION RESULTS: RECALL, PRECISION, EER AND THE REDUCTION OF ANALYZED IMAGES (IR) AND PERFROMED CLASSIFICATION STEPS (CR) IN PERCENT (COMPARED TO BS-V')

System	Rec	Prec	EER	IR	CR
Inertial Sensors (IS)	80	3	18	-	-
Basic System Vision (BS-V)	67	26	43	-	-
BS-V + opt HH (BS-V')	75	22	47	-	-
BS-V' + ROI	90	40	61	28	80
BS-V' + RLL	82	31	56	4	61
BS-V' + AH	78	29	55	19	66
BS-V' + ROI + RLL	87	48	63	38	84
BS-V' + ROI + AH	87	56	68	49	90
BS-V' + RLL + AH	82	46	64	36	85
BS-V' + AH + ROI + RLL	84	60	69	53	91
BS-V' + MoL	75	22	47	3	3
BS-V' + TF	75	22	47	21	14
BS-V' + NHM	75	22	46	42	35
BS-V' + OOM	81	35	54	0	51
No Infrastructure	78	30	55	32	70
BS-V' + ROI + AH + MoL + TF	86	57	68	56	91
BS-V' + ROI + AH + OOM	90	70	79	63	94

AH) and object operating mode features (OOM) are able to improve both recall and precision and reducing significantly the amount of analyzed images and performed classification steps in the same way. Comparing only location features, we can see that the combination of ROI and AH delivers the best result (a quite high recall of 87% and a EER of 68%). Time (TF) and movement features (MoL, NHM) are not able to increase recall nor the precision, but can still reduce significantly the number of processed images and classification steps. Finally we compare three main systems: First a system configuration which is only based on body-worn sensors and does not need any infrastructure (BS-V'+AH+MoL+TF), second a system without object operating mode monitoring (BS-V'+ROI+AH+MoL+TF) and finally the system setup reaching the highest EER (BS-V'+ROI+AH+OOM). For applications where it is difficult to instrument the environment our system is able to deliver a recall of 78% and a precision of 30%. Although the precision is quite low, our system provides a EER which is 37% higher than the EER that is reached when using only inertial sensors and statistical training (IS). Beside this, we can reduce the amount of classification steps by 70% and the analyzed images by 32%, which implicates a significant reduction of processing time. For applications where it is possible to use ceiling mounted cameras, we are able to increase again the recall by 8% and the precision by 27%, which results in a EER improvement of 13%. Beside this, the reduction of analyzed images can be improved by 24% and of classification steps by 21%. Finally if the instrumentation of electronic devices and the water tap is possible or even if intelligent devices are already available, our system is able to deliver again a significant improvement in both recall (finally 90%) and precision (finally 70%) which results in a final EER of 79%. Beside this the amount of analyzed images and classification steps was reduced again.

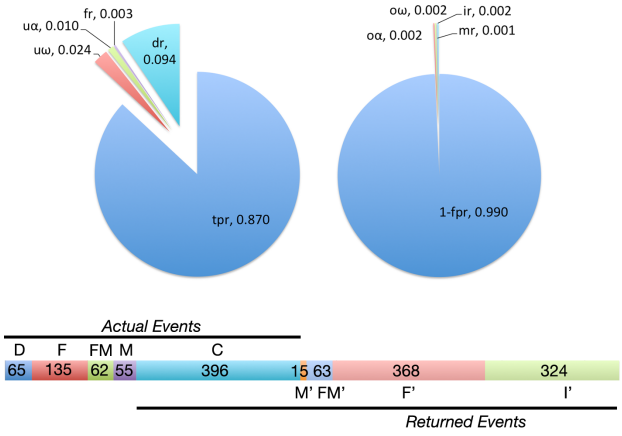

Figure 4. BS-V'+ROI+AH+OOM: Frame based evaluation - 2SET Metrics (top) and event based evaluation - EAD (bottom)

Having a look at the frame based evaluation (see Figure 4) we can see that 87% of the positive frames are correctly recognized and we have almost no fragmentations and underlays. Although we have a quite low precision regarding to our previous evaluation, we can see that the amount of insertions is vanishingly small in contrast to the amount of true negatives. Regarding the EAD evaluation we can see that 28% of all events are real insertions, in case of performed activities 9 % are real deletions, 56% are precise hits and nearly 35% are fragmented, merged or both.

VII. CONCLUSION

We have shown that a combination of a cheap proximity sensor, a inertial sensor based hand height estimation and a very simple vision algorithm applied to images delivered by a wrist mounted camera can spot subtle hand actions in a continuous data stream. Most importantly our approach avoids the need for a huge amount of statistically representative training data and instead relies on single "measurements" which can be easily performed in real life deployments.

The next step in our research is to investigate if the introduced approach can be even improved when combining it with a system based on inertial sensors and statistical training. Beside a higher recognition rate we expect also to get more detailed information about the performed activities. This means that the combination of both systems will be able to detect not only object interactions but also the specific activity which involves the object interaction.

REFERENCES

[1] O. Amft, H. Junker, and G. Tröster. Detection of eating and drinking arm gestures using inertial body-worn

sensors. In *Proceedings of the 9th IEEE International Symposium on Wearable Computers (ISWC)*, pages 160–163, 2005.

[2] Ling Bao and Stephen S. Intille. Activity recognition from user-annotated acceleration data. In *Proceedings of the 2nd International Conference on Pervasive Computing*, pages 1–17, April 2004.

[3] G. Bauer and P. Lukowicz. Developing a sub room level indoor location system for wide scale deployment in assisted living systems. In *Computers Helping People with Special Needs*, volume 5105 of *Lecture Notes in Computer Science*, pages 1057–1064. Springer Berlin / Heidelberg, 2008.

[4] G. Bauer, K. Stockinger, and P. Lukowicz. Recognizing the use-mode of kitchen appliances from their current consumption. In *Smart Sensing and Context*, volume 5741 of *Lecture Notes in Computer Science*, pages 163–176. Springer Berlin / Heidelberg, 2009.

[5] Ulf Blanke and Bernt Schiele. Daily routine recognition through activity spotting. In *4rd International Symposium on Location- and Context-Awareness (LoCA)*, 2009.

[6] A. Czabke, J. Neuhauser, and T.C. Lueth. Recognition of interactions with objects based on radio modules. 6 2010.

[7] T. Huynh and B. Schiele. Analyzing features for activity recognition. In *Proceedings of the ACM International Conference of the joint conference on Smart objects and ambient intelligence (EUSAI)*, Grenoble, France, 2005.

[8] Tâm Huynh, Mario Fritz, and Bernt Schiele. Discovery of Activity Patterns using Topic Models. In *Proceedings of the 10th ACM International Conference on Ubiquitous Computing (UbiComp)*, 2008.

[9] Alejandro Ibarz, Gerald Bauer, Roberto Casas, Alvaro Marco, and Paul Lukowicz. Design and evaluation of a sound based water flow measurement system. In *Smart Sensing and Context*, volume 5279 of *Lecture Notes in Computer Science*, pages 41–54. Springer Berlin / Heidelberg, 2008.

[10] Y. Ishiguro, A. Mujibiya, T. Miyaki, and J. Rekimoto. Aided eyes: eye activity sensing for daily life. In *Proceedings of the 1st Augmented Human International Conference*, page 25. ACM, 2010.

[11] Do-Un Jeong, Se-Jin Kim, and Wan-Young Chung. Classification of posture and movement using a 3-axis accelerometer. In *Proceedings of the 2007 International Conference on Convergence Information Technology*, ICCIT '07, pages 837–844, Washington, DC, USA, 2007. IEEE Computer Society.

[12] G. LeBellego, N. Noury, G. Virone, M. Mousseau, and J. Demongeot. A model for the measurement of patient activity in a hospital suite. *Trans. Info. Tech. Biomed.*, 10(1):92–99, January 2006.

[13] Takuya Maekawa, Yutaka Yanagisawa, Yasue Kishino, Katsuhiko Ishiguro, Koji Kamei, Yasushi Sakurai, and Takeshi Okadome. Object-based activity recognition with heterogeneous sensors on wrist. In *Pervasive Computing*, volume 6030 of *Lecture Notes in Computer Science*, pages 246–264. Springer Berlin Heidelberg, 2010.

[14] EmmanuelMunguia Tapia, StephenS. Intille, and Kent Larson. Activity recognition in the home using simple and ubiquitous sensors. In *Pervasive Computing*, volume 3001 of *Lecture Notes in Computer Science*, pages 158–175. Springer Berlin Heidelberg, 2004.

[15] T. Toyama, T. Kieninger, F. Shafait, and A. Dengel. Gaze guided object recognition using a head-mounted eye tracker. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 91–98. ACM, 2012.

[16] K. Van Laerhoven and O. Cakmakci. What shall we teach our pants. In *Proceedings of the 4th IEEE International Symposium on Wearable Computers (ISWC)*, pages 77–83. Citeseer, 2000.

[17] Jamie A. Ward, Paul Lukowicz, and Hans W. Gellersen. Performance metrics for activity recognition. *ACM Trans. Intell. Syst. Technol.*, 2(1):6:1–6:23, January 2011.

[18] Andreas Zinnen, Kristof Van Laerhoven, and Bernt Schiele. Toward recognition of short and non-repetitive activities from wearable sensors. In *AmI*, volume 4794 of *Lecture Notes in Computer Science*, pages 142–158. Springer, Springer, 2007.