Scalable Recognition of Daily Activities with Wearable Sensors

Tâm Huỳnh, Ulf Blanke and Bernt Schiele

Computer Science Department
TU Darmstadt, Germany
{huynh, blanke, schiele}@mis.tu-darmstadt.de

Abstract. High-level and longer-term activity recognition has great potentials in areas such as medical diagnosis and human behavior modeling. So far however, activity recognition research has mostly focused on low-level and short-term activities. This paper therefore makes a first step towards recognition of high-level activities as they occur in daily life. For this we record a realistic 10h data set and analyze the performance of four different algorithms for the recognition of both low- and high-level activities. Here we focus on simple features and computationally efficient algorithms as this facilitates the embedding and deployment of the approach in real-world scenarios. While preliminary, the experimental results suggest that the recognition of high-level activities can be achieved with the same algorithms as the recognition of low-level activities.

1 Introduction

Activity recognition has been an active area of research in recent years due to its potential and usefulness for context-aware computing. Current approaches typically rely on state-of-the-art machine learning ranging from unsupervised to supervised techniques and from discriminant to generative models. Most research however has focused on low-level and short-term activities. While this focus has advanced the state-of-the-art significantly we strongly believe that activity recognition should move forward to address the important and challenging area of longer-term and high-level activity recognition. In many applications ranging from medical diagnosis over elderly care to modeling of human behavior, the analysis and recognition of high-level activities is an important component.

There are various reasons why only a few researchers have worked on longer-term, complex and high-level activities (with some notable exceptions as discussed in Section 2). For example it is often argued that the recognition of low-level activities is a prerequisite to recognize more complex and high-level activities. Besides being tedious and time-consuming, the recording of high-level activities is a non-trivial task, as the data should be as realistic and representative as possible. So fundamental problems such as the inherent difficulties and the large variability as well as more practical reasons seem to have prevented most researchers to address the recognition of complex and high-level activities.

The explicit goal of our research is to enable the recognition of longer-term and high-level activities. Therefore, an essential first step is to record an interesting and realistic dataset of high-level activities. As we are interested in long-term activities it is essential to use long-term recordings which is why this papers uses over 10h worth of data. The paper then compares four algorithms both for the recognition of low-level activities as well as high-level activities. For each of the algorithms, we analyze and discuss different parameters such as feature length and sensor placement. The results suggest that the recognition of high-level activities may be achievable with the same algorithms as for low-level activities. In particular, our results indicate that recognition of high-level activities can be achieved using features computed from raw sensor data alone, without building up any intermediate representation such as a grammar of low-level activities.

Let us briefly define – for the purpose of this paper – the difference between low-level and high-level activities. Low-level activities are e.g. *walking, sitting, standing, hoovering, eating, washing dishes*, etc which typically last between 10s of seconds to several minutes. High-level activities, on the other hand, are longer-term as e.g. *cleaning the house*, which will typically last more than 10s of minutes and could last as long as a few hours.

The main contributions of the paper are as follows. First, the results of our experiments suggest that today's activity recognition algorithms are quite capable to address the problem of high-level activity recognition. Second, we record and provide an interesting and realistic dataset of high-level activities which we plan to make publicly available upon publication of this paper. Third, we analyze and compare different algorithms for the recognition of low-level and high-level activities. Fourth, we systematically analyze important parameters such as sensor placement, feature length and classification window.

The paper is structured as follows: In the next section we will put our work into context by discussing related work. In Section 3, we introduce the dataset and hardware for our experiments. Section 4 presents the algorithms we use for recognition of both high- and low-level activities. Sections 5 and 6 report on the results for low- and high-level activities, respectively. Section 7 presents the summary and conclusion.

2 Related Work

Current research in activity recognition from wearable sensors covers a wide range of topics, with research groups focusing on topics such as the recognition of activities of daily living (ADLs) in the context of healthcare and elderly care (e.g. [1]), automated discovery of activity primitives in unlabeled data (e.g. [2]), semi- or unsupervised learning of activities (e.g. [3, 4]), or the combination of several sensor modalities to improve recognition performance (e.g. [5, 6]). The majority of this work is concerned with single activities over relatively short timescales, ranging from limb movements in dumbbell exercises [2] over postures and modes of ambulation such as *sitting, standing* and *walking* [7, 8], to household activities such as *making tea, dusting, cleaning the windows* or *taking a*

shower [6, 9]. To our knowledge, little work has been done in using wearable sensors to recognize activities on larger time scales, i.e. by recognizing higher-level scenes such as *cleaning the house* or *going shopping*. A notable exception is the work by Clarkson et al. [10], who used wearable vision and audio sensors to recognize scenes such as a user visiting a supermarket or a video store. However, since cameras and microphones are considered intrusive by many people, such an approach is unlikely to be adopted in everyday life. There has been work in identifying daily routines in the lives of users (e.g. [11]) or inferring a user's high-level intentions during his daily movements through urban environments (e.g., [12–14]). However, these works mainly focus on the location of the user or have a different understanding of the term 'high-level', more referring to a user's abstract goals in terms of traveling destinations than to a collection of related low-level activities. On a smaller scale, [6] proposed to break down activities such as *cleaning the windows* into small movements called *actions*, such as *wipe horizontally* and *wipe vertically*. In this work we follow a different approach, by summarizing a collection of activities into scenes measured in hours rather than in minutes.

3 Experimental Setup

An important first step towards the recognition of high-level activities is a realistic and representative recording of sensor-data. We formulated four requirements and considerations as the basis of our data recording. First, as the primary aim is the recognition of high-level activities, we explicitly started with the recording of such activities and later defined, named and annotated those low-level activities that occurred and were performed during these high-level activities. As we will see below, this leads to quite a different set of low-level activities than one may obtain when starting from low-level activities. Second, the recording should be as realistic as possible so that the activities should be performed "in the field" – that is in an unconstrained and natural setting – and not in a laboratory or staged setting. Third, the usefulness and the usability of high-level activity recognition strongly depends on the price and form-factor of the final device. Therefore we decided to keep the algorithms, features and the sensor-platform as simple and power-efficient as possible so that the embedding into a simple self-contained device is feasible in the future. Forth, we decided to start with the recording of data for a single user, as our primary aim in this paper is to analyze and show the feasibility of high-level activity recognition first. Even though that might seem like a limitation, we rather expect that the execution of high-level activities varies greatly between individuals so that one might need to use a personalized device. If this holds true or one can enable person-independent high-level activity recognition remains an open research question and is beyond the scope of this paper.

One requirement formulated above was to base our recognition on simple sensors and easy-to-compute features which is why we decided to use the mean and variance of acceleration signals. Accelerometers are especially appealing in

this context, since they are cheap and can be increasingly found in everyday objects such as mobile phones, cameras, wrist watches and even shoes. The use of simple features for recognition would allow the computation to take place online on a miniature mobile device without draining the battery or slowing down other applications. Computing the features on the device and discarding the raw signals can also help to save memory and allow for longer recordings.

Dataset. During the recordings the user was wearing three sensors. One sensor was attached to the right wrist, one to the righthand side of the hip, and one to the right thigh, as illustrated in Figure 3(a). The ground truth labels were mainly added and edited offline, using a separate video recording (from a passively mounted video-camera used during the *housework* and *morning* scenes) and some optional online annotations from a PDA.

The dataset consists of three different high-level activities or *scenes* performed by one user. The first scene consists of a typical morning routine one might perform before going to work, which, for one of the recordings, looked as follows (see Figure 1 for the corresponding ground truth annotation). After some time of sleeping, the user gets up, walks to the bathroom, uses the toilet and brushes his teeth. After having breakfast, he leaves the house and drives to work by car. The second scene is a shopping scenario which might look as follows: after working at the computer for some time, the user walks to his car and drives to a nearby shopping center, buys groceries and heads back in his car. In the third scene, the user does some housework after getting up. He might first brush his teeth and have some breakfast, may then wash the dishes, hoover his apartment and iron some clothes, and eventually walk out of the house.

Each scene was recorded four times, on different days and in a natural environment, i.e. at the user's home and in a nearby supermarket. The scenes were loosely defined by the fact that each activity should at least occur once in each instance. The length of the scenes varies between 40 and 80 minutes; the total length of the data is 621 minutes. Figure 1 shows the ground truth for one instance of each scene, and Figure 2 gives an overview of all activities. The scenes consist of 15 different activities (plus one garbage class for unlabeled data), some of which are shared between two or three scenes. For evaluation, we created four sets, each consisting of three concatenated scenes. We used these sets to perform a 4-fold leave-one-out crossvalidation on the data.

Hardware. Figure 3(b) shows the sensor platform that was used for recording the data for our experiments [15]. It features a 2D accelerometer (ADXL202JE) and nine binary tilt switches for sensing motion and orientation of the user. The sensor board is stacked onto a BSN node [16] with 512 kb of EEPROM storage for logging sensor data, followed by a third board for the power supply.

Feature Computation. During recordings, the platform stores all sensor data on the EEPROM storage, from which it can later be retrieved via an rs232 connection. As we aimed for recordings of several hours, the limiting factor for our experiments was the size of the 512 kb on-board memory rather than battery

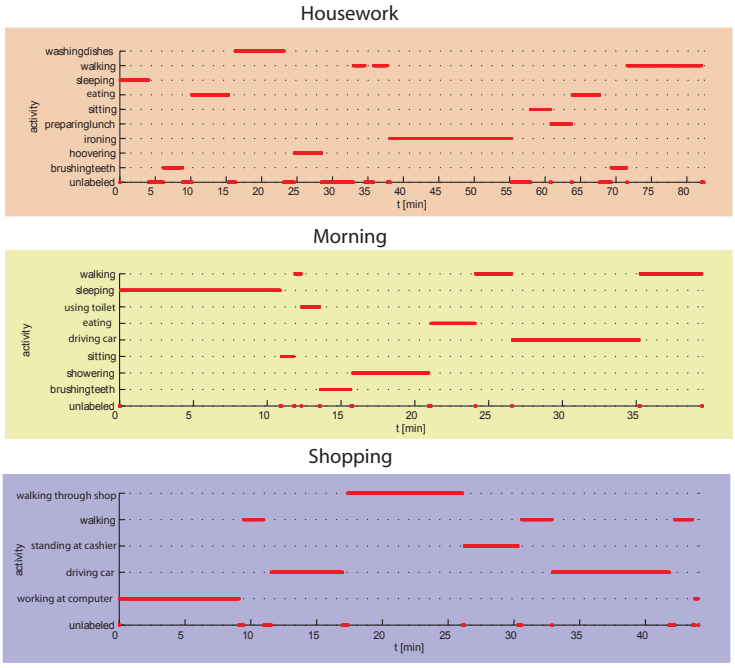


Fig. 1. Ground truth for recordings of the three scenes *Housework*, *Morning* and *Shopping*. Each scene was performed four times by the user, here we show only one instance of each scene.

lifetime. To save memory, we compute and store only the mean and variance of the acceleration signal at 2 Hz and discard the raw (80 Hz) acceleration data. This allows us to record about five hours of sensor data on the chip. The next generation of the platform will have a larger on-board memory and allow for recordings of several days or even weeks.

4 Algorithms

We use four different approaches for recognition of activities – three of them are based on a discrete representation that we obtain by clustering the sensor data, and one approach is based on training HMMs on continuous data. All approaches have in common that they use the mean and variance of the acceleration signal over a sliding window as the underlying features. These features are cheap to compute and are known to yield high recognition rates in settings comparable to ours (e.g. [8, 17–19]).

Related work has shown that it is possible to recognize movements or activities based on low dimensional models learned in a semi- or unsupervised fashion (e.g., [2, 19]). Such models can also be thought of as an alphabet of symbols, a vocabulary in which activities are formulated as 'sentences'. Compositions of

Highlevel Activities	*Lowlevel Activities*	
a Preparing for Work	1 (unlabeled)	9 walking [a, b]
b Going Shopping	2 brushing teeth [a, c]	10 working at computer [b]
c Doing Housework	3 taking a shower [a]	11 waiting in line in a shop [b]
	4 sitting [a]	12 strolling through a shop [b]
	5 driving car [a, b]	13 hoovering [c]
	6 eating at table [a,c]	14 ironing [c]
	7 using the toilet [a]	15 preparing lunch [c]
	8 sleeping [a]	16 washing the dishes [c]

Fig. 2. Overview of the low- and high-level activities in the recorded dataset. Each high-level activity consists of a set of low-level activities, as indicated in brackets.

(a) (b)

Fig. 3. Left: User wearing sensors on wrist, hip and thigh. Right: The sensor platform, consisting of the power supply (bottom), the BSN node for logging (middle) and the sensor board (top).

such sentences could later serve as a tool for recognizing more abstract and high-level behavior. The first three of the following approaches are inspired by this idea, but as we do not assume that human motion follows a strict grammar, we only consider the occurrences of symbols over intervals, without modeling their temporal order. We use k-means clustering as a simple yet effective unsupervised method to map features to a set of discrete symbols, i.e. to one of the k cluster centers. We represent each feature by the closest cluster center. As a result, the input data is transformed into a one-dimensional sequence of cluster assignments. Based on this representation, we employ three different learning methods which we describe in the following. The fourth method is based on HMMs and uses a vector of mean and variance values as features. Figure 4 illustrates the different representations we use for recognition.

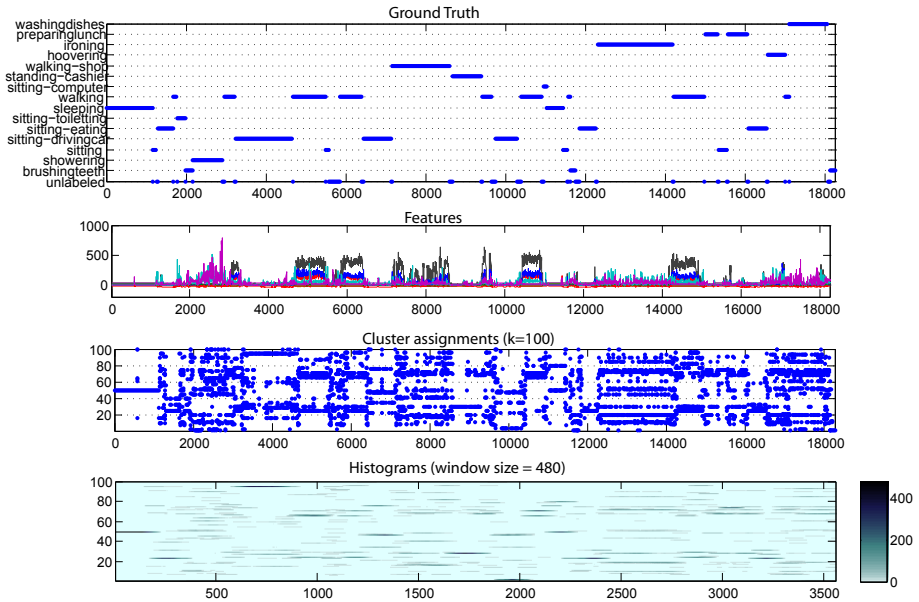


Fig. 4. Examples of the different representations used for recognition. From top to bottom: ground truth; features (mean & variance over 4 sec); cluster assignments (each feature is assigned to one of k=100 clusters); histograms of cluster assignments (over windows of 480 samples).

K-means. As a baseline method, we label each cluster with the activity that occurs most often among the training samples belonging to the cluster. Classification is then performed by assigning to each test sample the label of the closest cluster center. During experiments we vary the size of k and the length of the window over which the features are computed.

Occurrence Statistics + NN. In this approach, rather than using individual symbols as features, we compute histograms of cluster assignments over a sliding window of the training sequence. Each histogram is labeled with the activity that occurs most often in the window of samples that it covers. For evaluation, we perform a nearest neighbor (NN) classification on the histograms computed from a test sequence.

Occurrence Statistics + SVM. This approach is also based on histograms of cluster assignments. However, instead of using a nearest neighbor classifier, we train a support vector machine (SVM) using the histograms as features.

HMMs. The fourth approach is based on Hidden Markov Models (HMMs). HMMs belong to the class of generative statistical signal models, and they have been successfully used in activity recognition tasks before (e.g. [20, 10, 21]). They

lend themselves to a hierarchical classifier design, which makes them interesting candidates for modelling activities on different levels of abstraction.

As for the first three approaches, we use the mean and variance of the acceleration signal over a sliding window as features. We then partition the data into N equal parts and train a separate HMM on each part. We use left-right models with one gaussian per state, and we vary the number of states in our experiments. In order to assign activity labels to the models, we use a sliding window over the features as observation sequence, and compute the likelihood of the window for each of the N models. The model with the highest likelihood is then assigned the label of the activity that occurs most often in the window. Classification is performed similarly, i.e. by computing the likelihood of each model over a sliding window starting at a certain sample, and subsequently assigning to the sample the label of the model with the highest likelihood.

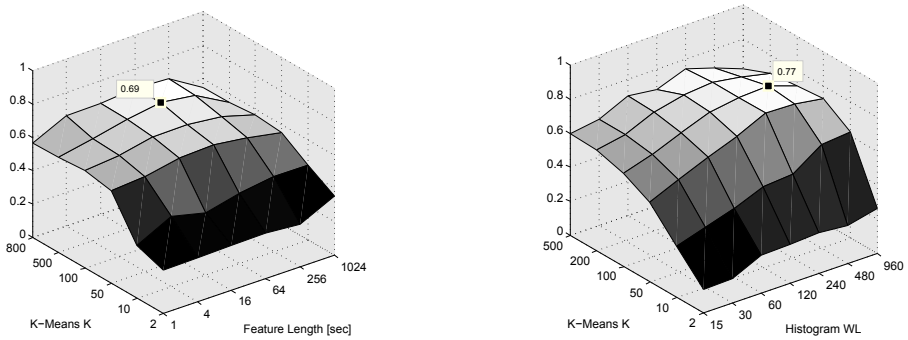
5 Low-level Activities

In this section we report on the performance of our proposed approaches with respect to the fifteen low-level activities listed in Figure 2. As mentioned earlier, the definition of those low-level activities came after the recording of the high-level activities. That way, a somewhat obvious but important observation is that the definition of low-level activities is not as well-defined as one might expect. E.g., for the following activities, it is not clear if they belong to the same or to different low-level activities: *walking down a corridor* vs. *walking in a supermarket while collecting items*; *sitting in a car* vs. *sitting at a table while eating* vs. *sitting on the toilet* vs. *sitting at a desk and working on a computer*; etc. It should be clear that this is not simply a question of a hierarchical and temporal decomposition of concurrent activities but that this is rather an inherent difficulty linked to the context of the particular activity (e.g. *sitting on the toilet* vs. *sitting at a table*). So we decided to define the low-level activities within each high-level activity as they occurred within the context of the high-level activity. That way we have a range of activities which occur across multiple high-level activities such as *walking, eating at table* and *brushing teeth* and others which are more specific such as *driving a car* or *strolling through a shop*.

Based on these definitions of low-level activities, this section compares he recognition performance of our four approaches. For each of the algorithms we also identify and discuss suitable parameters such as the number of clusters, the length of the feature window, and also appropriate on-body locations for the sensors.

K-means. Figure 5(a) shows the accuracy[1] for different numbers k of clusters and different window lengths for the features. One can observe that values of k below 50 have a negative impact on the recognition performance. For values of $k >= 50$, accuracy lies roughly between 60 and 70%. The best result of 69,4% is

[1] we use the term *accuracy* to refer to the number of correctly classified samples divided by the number of all samples



(a) (b)

Fig. 5. Accuracy of classification for low-level activities; using assignments to cluster centers as features (left) vs. using histograms of such assignments in combination with nearest neighbor classification (right).

		Classified Activity																	
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**Sum**	**Recall**
Ground Truth	**1 unlabeled**	**1314**	148	188	64	567	136	81	103	540	85	10	428	317	993	707	158	5839	**22.5%**
	2 brush teeth	146	**1258**	310	0	0	0	0	0	9	0	16	0	71	302	2	51	2165	**58.1%**
	3 shower	83	249	**1710**	0	0	13	17	0	58	0	0	47	54	270	40	70	2611	**65.5%**
	4 sit	287	4	6	**684**	424	168	267	5	68	19	0	49	7	26	19	0	2033	**33.6%**
	5 drive car	334	0	0	343	**9743**	301	84	26	62	0	0	73	0	0	0	0	10966	**88.8%**
	6 eat	192	5	21	127	938	**4253**	26	11	42	0	0	25	2	3	21	13	5679	**74.9%**
	7 use toilet	83	14	46	41	324	106	**224**	7	14	16	0	0	12	53	4	0	944	**23.7%**
	8 sleep	260	14	21	45	116	34	0	**7016**	111	36	0	55	0	0	10	22	7740	**90.6%**
	9 walk	614	12	105	29	66	0	7	0	**8988**	0	0	1285	139	153	32	40	11470	**78.4%**
	10 work at comp.	99	0	3	22	52	35	15	36	26	**1325**	0	24	9	21	21	0	1688	**78.5%**
	11 stand at cashier	14	0	0	0	0	0	0	0	0	0	**798**	717	23	145	92	15	1804	**44.2%**
	12 walk in shop	193	14	37	7	2	0	0	0	836	0	297	**3260**	109	201	342	14	5312	**61.4%**
	13 hoover	74	44	74	0	0	0	0	0	128	0	0	135	**785**	456	66	149	1911	**41.1%**
	14 iron	122	76	155	0	0	0	0	0	38	0	162	53	267	**7009**	438	263	8583	**81.7%**
	15 prep. lunch	331	4	2	0	0	20	0	0	14	0	37	349	49	499	**731**	95	2131	**34.3%**
	16 wash dishes	240	29	54	13	0	0	0	0	11	0	3	37	23	350	255	**2554**	3569	**71.6%**
	Sum	4386	1871	2732	1375	12232	5066	721	7204	10945	1481	1323	6537	1867	10481	2780	3444	74445	
	Precision	**30.0%**	**67.2%**	**62.6%**	**49.7%**	**79.7%**	**84.0%**	**31.1%**	**97.4%**	**82.1%**	**89.5%**	**60.3%**	**49.9%**	**42.0%**	**66.9%**	**26.3%**	**74.2%**		

Fig. 6. Aggregate confusion matrix for the best parameter combination when using k-means cluster centers as features. $k = 500$, mean & var computed over 64 seconds, shift = 0.5 seconds. Overall accuracy is 69%.

obtained for $k = 500$ and a feature length of 64 seconds. Surprisingly, the best results are obtained for relatively long window lengths. Lengths between 16 and 256 seconds perform best, and there is a visible drop in performance for shorter and longer window lengths. Figure 6 shows the confusion matrix for the best parameter combination. One can clearly see that the recognition performance varies strongly between the different activities. Seven of the 15 activities have recall or precision values above 70%, the best being *sleeping* (97.4/90.6), *working at the computer* (89.9/78.5), *walking* (82.1/78.4) and *driving car* (79.7/88.8). During four activities the user was sitting (*sitting, driving car, eating at table, using the toilet*), and from Figure 6 one can see that these activities are often confused with each other during classification.

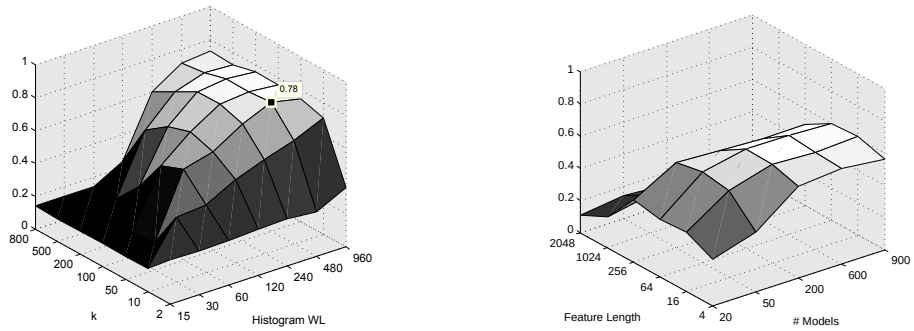
Ground Truth	Classified Activity																	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Sum	Recall
1 unlabeled		**79**	79	24	0	7	4	34	50	53	0	0	0	10	208	11	66	625	**12.6%**
2 brush teeth		73	**142**	86	0	0	8	8	1	0	0	0	3	0	0	0	59	380	**37.4%**
3 shower		0	0	**558**	0	0	0	0	0	0	0	0	0	0	0	0	3	561	**99.5%**
4 sit		0	0	0	**0**	67	0	68	0	26	0	0	0	0	0	27	0	188	**0.0%**
5 drive car		0	0	0	0	**2134**	0	0	0	102	0	0	54	0	0	0	0	2290	**93.2%**
6 eat		0	10	13	0	104	**1033**	25	0	0	0	0	47	5	0	11	23	1271	**81.3%**
7 use toilet		0	81	5	0	71	28	**26**	10	13	0	0	0	0	0	0	0	234	**11.1%**
8 sleep		15	26	0	0	0	4	7	**1498**	7	0	0	0	0	0	0	0	1557	**96.2%**
9 walk		46	0	6	0	90	0	17	7	**1632**	18	0	303	10	69	0	3	2201	**74.1%**
10 work at comp.		0	0	0	0	127	0	0	0	17	**180**	0	0	0	0	0	0	324	**55.6%**
11 stand at cashier		0	0	0	0	0	0	0	0	3	0	**125**	120	0	0	139	0	387	**32.3%**
12 walk in shop		0	0	0	0	23	0	0	0	52	0	60	**886**	0	0	75	0	1096	**80.8%**
13 hoover		10	0	0	0	0	0	0	0	0	0	0	0	**365**	8	10	21	414	**88.2%**
14 iron		0	0	0	0	0	10	0	0	10	0	0	0	0	**1676**	45	10	1751	**95.7%**
15 prep. lunch		0	0	8	68	0	17	30	0	5	0	30	53	13	14	**156**	109	503	**31.0%**
16 wash dishes		0	0	12	0	0	17	0	0	6	0	0	14	6	0	0	**715**	770	**92.9%**
	Sum	223	338	712	68	2623	1121	215	1566	1926	198	215	1480	409	1975	474	1009	14552	
	Precision	**35.4%**	**42.0%**	**78.4%**	**0.0%**	**81.4%**	**92.1%**	**12.1%**	**95.7%**	**84.7%**	**90.9%**	**58.1%**	**59.9%**	**89.2%**	**84.9%**	**32.9%**	**70.9%**		

Fig. 7. Aggregate confusion matrix for the best parameter combination when using histograms of cluster centers as features. $k = 100$, histogram windows over 480 features (about 4 min.) shifted by 5 features each, mean & var computed over 4 sec., shift = 0.5 seconds. Overall accuracy is 77%.

Occurrence Statistics + NN. Figure 5(b) shows the recognition results for the histogram-based approach combined with a nearest neighbor classifier. We vary the number of clusters and the length of the histogram windows (the windows are always shifted by 5 features at a time). The underlying mean and variance features are computed over windows of 4 seconds with a shift of 0.5 seconds (in contrast to the k-means approach, we observed that small feature windows performed better here). The highest accuracy of 77% is obtained for $k = 100$ and a histogram window of 480 samples, covering about 4 minutes of data. For larger histogram windows the accuracy visibly decreases. Similarly to the k-means results, values of k below 50 lead to a sharp drop in performance, implying that too much information is lost from the discretization. Figure 7 shows the confusion matrix for the best parameter settings. Except for the activities *taking a shower*, *sitting*, *using the toilet* and *washing the dishes*, the precision increases for all activities compared to the previous approach. Notably, the confusion between the activities *ironing* and *hoovering* is much lower in this approach. The overall gain in accuracy of 8% indicates that the use of histograms of symbols rather than individual symbols does indeed help to improve recognition performance.

Occurrence Statistics + SVM. When using an SVM for classification in combination with the histogram features, the recognition results can be slightly improved compared to the nearest neighbor approach. Figure 8(a) shows the accuracy for different values of k and different window lengths for the histograms. The best result of 78% is obtained for $k = 50$ and a histogram window of 480 samples, covering about 4 minutes of data. One can observe that accuracy decreases with higher number of clusters and smaller window lengths. For window lengths between 240 and 960 samples, corresponding to about 2 to 8 minutes of data, and values of k between 50 and 200, we obtain the highest accuracies.

HMMs. Figure 8(b) shows recognition results for the HMM approach. We vary the feature length and the number of models N; in this particular example,



(a) (b)

Fig. 8. Accuracy of classification for low-level activities; using histograms of cluster assignments in combination with an SVM (left) vs. using HMMs (right).

Ground Truth	Classified Activity 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Sum	Recall
1 unlabeled	**288**	337	654	849	323	497	533	49	879	458	65	833	103	485	151	98	6602	**4.4%**
2 brush teeth	0	**1565**	32	77	21	32	46	0	20	32	0	0	0	159	183	0	2167	**72.2%**
3 hoover	0	75	**1070**	523	0	33	32	0	120	50	0	0	0	0	0	0	1903	**56.2%**
4 iron	0	0	182	**6732**	1211	41	0	0	117	166	0	0	0	0	130	0	8579	**78.5%**
5 prep. lunch	0	0	0	365	**927**	105	27	0	131	162	66	35	130	33	54	33	2068	**44.8%**
6 sit	10	81	0	61	142	**577**	84	13	273	212	33	257	0	0	0	196	1939	**29.8%**
7 eat	0	0	20	43	70	80	**3615**	0	130	252	0	1327	0	33	19	32	5621	**64.3%**
8 sleep	533	16	0	0	11	283	29	**6638**	56	0	33	127	0	0	0	65	7791	**85.2%**
9 walk	33	130	196	145	7	463	310	0	**8399**	171	0	493	13	903	0	33	11296	**74.4%**
10 wash dishes	0	9	0	167	98	69	206	0	2	**2740**	0	0	0	0	262	0	3553	**77.1%**
11 work at comp.	0	0	0	0	0	65	24	0	178	0	**1307**	33	0	86	0	0	1693	**77.2%**
12 drive car	0	0	0	0	0	99	524	0	440	0	33	**9670**	0	33	0	98	10897	**88.7%**
13 stand at	99	0	0	0	200	0	0	0	0	0	0	0	**1285**	212	0	0	1796	**71.5%**
14 walk in shop	0	98	0	0	197	66	0	0	862	0	0	68	429	**3581**	0	0	5301	**67.6%**
15 shower	0	254	0	205	0	0	0	0	0	130	0	0	0	0	**1982**	0	2571	**77.1%**
16 use toilet	20	16	0	25	0	206	328	0	28	2	0	295	0	0	0	0	920	**0.0%**
Sum	983	2581	2154	9192	3207	2616	5758	6700	11635	4375	1537	13138	1960	5525	2781	555	74697	
Precision	**29.3%**	**60.6%**	**49.7%**	**73.2%**	**28.9%**	**22.1%**	**62.8%**	**99.1%**	**72.2%**	**62.6%**	**85.0%**	**73.6%**	**65.6%**	**64.8%**	**71.3%**	**0.0%**		

Fig. 9. Aggregate confusion matrix for the best parameter combination when using the HMM-based approach. The parameters were: window length for features = 64 sec., 200 models, 32 states per model, observation length = 16. Overall accuracy is 67.4%.

the number of states is fixed to 8, and the observation window for classification covers 16 samples. The number of models N directly affects the length of data that each HMM models, since the data is equally partitioned into N parts. Thus, N is inversely related to the length of the histogram windows of the previous approaches. From the plot one can observe that using less than 200 models (i.e. each model sees about 2.5 min of data or more) leads to a visible decrease in performance. We obtained the best result of 67% for $N = 200$ models and a feature length of 64 sec, an observation length of 16 and models with 32 states. When varying the number of states we found that they only marginally effected the results. Figure 9 shows the confusion matrix for the best parameter combination. Overall, results of the HMM approach suggest that the temporal aspect – at least for the features we employed – is not dominant enough to allow for higher recognition rates.

Sensor placement. The results so far were based on the data of all three sensors the user was wearing on wrist, hip and thigh. It turns out that using only subsets of these sensors for recognition reveals some interesting relations between the placement of sensors and the recognition of individual activities. For instance, we found that the overall accuracy of the k-means approach slightly improved from 69 to 70% when we used only two sensors, namely the sensors on wrist and thigh. These results are consistent with the findings from [8], who also found that when using only two sensor locations, wrist and thigh are the most suitable locations. Using these locations even leads to better results when recognizing the activities *brushing teeth*, *driving car*, *preparing lunch* and *washing dishes*. When only using the wrist sensor, performance for *brushing teeth* and *taking a shower* improves, likely because these activities are mainly characterized by hand and arm movements. For *sleeping* and *walking*, using only the hip sensor already yields precision and recall values up to 95%.

5.1 Discussion

Figure 10 shows a summary table comparing the best results of the four approaches. Generally, the approach *Occurrence Statistics + SVM* achieves the highest accuracy of 79.1%. For most activities, the use of histograms instead of single cluster assignments as features leads to better precision and recall values. However, there are two stationary (*sitting, using the toilet*) and two dynamic activities (*brushing teeth, walking*) in which the use of single cluster assignments yields higher results in either precision, recall or both. The HMM approach achieves the lowest accuracy of 67.4%, slightly less than the *k-means* approach. In summary, we conclude that using histograms of symbols as features and combining them with a strong classifier is a promising and competitive approach for recognizing the type of daily activities we recorded in our study.

It is worth noting that the overall recognition scores seem low compared to the published state-of-the-art. However, in contrast to most other recordings and as discussed above, we explicitly defined the low-level activities after the recording of the high-level activities, and therefore both the larger variability within single low-level activities (such as *walking*) and the high similarity between different low-activities (such as *walking* and *walking through shop*) pose a more challenging recognition problem than is usually addressed.

6 High-level Activities

In this section we report on how well our proposed approaches can deal with the recognition of high-level scenes comprising a collection of low-level activities. More specifically, we evaluate how well our algorithms can classify the three different scenes *Morning*, *Housework*, and *Shopping*. Each scene has a length of at least 40 minutes and consists of at least six different activities. The evaluation was performed in the same fashion as for the low-level activities: we constructed four datasets, each containing one instance of each of the three scenes, and then performed a leave-one-out crossvalidation.

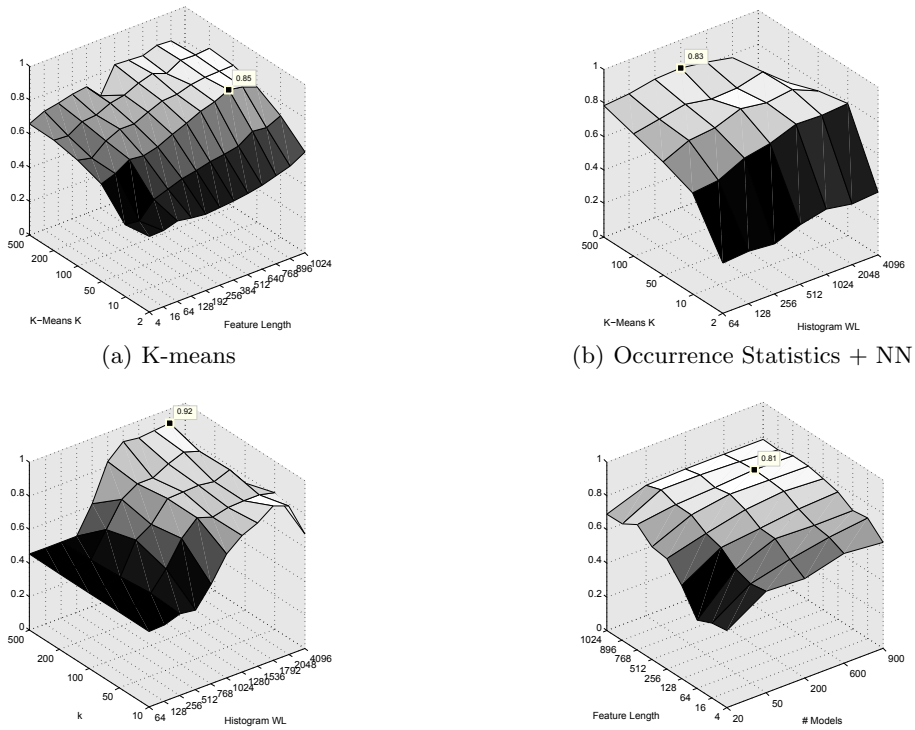
	K-means		Occ./NN		Occ./SVM		HMM	
Activity	p	r	p	r	p	r	p	r
(unlabeled)	30,0	**22,5**	**35,4**	12,6	7,9	3,0	29,3	4,4
brush teeth	**67,2**	58,1	42,0	37,4	23,0	21,1	60,6	**72,2**
shower	62,6	65,5	78,4	**99,5**	**86,7**	91,8	71,3	77,1
sit	**49,7**	**33,6**	0,0	0,0	0,0	0,0	22,1	29,8
drive car	79,7	88,8	81,4	93,2	**86,9**	**95,4**	73,6	88,7
eat	84,0	74,9	**92,1**	81,3	82,3	**87,3**	62,8	64,3
use toilet	**31,1**	**23,7**	12,1	11,1	15,0	9,4	0,0	0,0
sleep	97,4	90,6	95,7	96,2	91,2	**97,2**	**99,1**	85,2
walk	82,1	**78,4**	**84,7**	74,1	79,5	77,6	72,2	74,4
work at computer	89,5	78,5	90,9	55,6	**93,3**	**94,8**	85,0	77,2
stand at cashier	60,3	44,2	58,1	32,3	**75,9**	47,3	65,6	**71,5**
walk in shop	49,9	61,4	59,9	**80,8**	**70,7**	80,1	64,8	67,6
hoover	42,0	41,1	89,2	**88,2**	**98,3**	82,6	49,7	56,2
iron	66,9	81,7	84,9	95,7	**89,0**	**95,9**	73,2	78,5
prep. lunch	26,3	34,3	32,9	31,0	**45,7**	**54,3**	28,9	44,8
wash dishes	74,2	71,6	70,9	**92,9**	**79,0**	89,9	62,6	77,1
Mean	62,0	59,3	63,0	61,4	**64,0**	**64,2**	57,5	60,6
Accuracy	69,4		77,0		**79,1**		67,4	

Fig. 10. Summary of the results for low-level activities. Each column shows the precision (p) and recall (r) values for each activity, as well as the accuracy, i.e. the number of correctly classified samples divided by all samples. The highest values in each row are highlighted.

K-means. Figure 11(a) shows the accuracy for different numbers of clusters and different window lengths for computing mean and variance of the signal. As for the low-level activities, one can observe that for values of k below 50 performance decreases rapidly. In terms of feature windows, there is a visible tendency that longer window lengths lead to a better performance. For the parameter values that we sampled, the best result of 84.9% was obtained for $k = 50$ and a feature window of 768 sec., i.e. about 13 min. (We comment on the feature length below in the paragraph 'Sensor Placement'.) The confusion matrix for this configuration is shown in Figure 12 (upper left). Precision and recall range between 74 and 94%.

Occurrence Statistics + NN. In this experiment, as for the low-level activities, we vary the number of clusters and the length of the histogram. The results can be seen in Figure 11(b). The mean and variance features are computed over 4 sec. windows with a shift of 1 second. The best results are obtained for values of k between 50 and 500, and histogram windows between 512 and 2048 samples, i.e. between about 8 and 32 minutes. Figure 12 (upper right) shows the confusion matrix for $k = 500$ and a histogram window of 512 samples; the accuracy for this run was 83.4%, which is slightly lower than for the k-means approach. In terms of precision and confusion there is no clear difference to the k-means approach. However, the results improve substantially when using an SVM for classification instead of a nearest neighbor classifier, as is described in the next section.

Occurrence Statistics + SVM. Figure 11(c) shows the accuracy for different values of k and different window lengths for the histograms when using an SVM as



(a) K-means

(b) Occurrence Statistics + NN

(c) Occurrence Statistics + SVM

(d) HMM

Fig. 11. Accuracy of classification for high-level activities.

classifier. The best results are obtained for histogram windows between 1280 and 2048 samples, i.e. between 20 and 32 min. Interestingly, the number of clusters for discretization only has a minimal influence on the recognition performance, the dominating parameter is the length of the histogram window. Even when using only $k = 10$ clusters, the accuracy stays above 90%. Figure 12 (lower left) shows the confusion matrix for the best result of 91.8% accuracy, which is an improvement of about 7% compared to using the nearest neighbor classifier as described in the previous paragraph.

HMMs. Figure 11(d) shows the recognition results for the HMM approach. As for the low-level activities, we vary the feature length and the number of models N. The number of states is fixed to $s = 2$ (we did vary the number of states but found only small changes in performance), and the length of the observation window for each HMM is set to 16 samples. From the figure one can observe that values of N below 200 lead to a decrease in performance. The best results of slightly above 80% are obtained for feature lengths above 256 seconds (4 min) and $N = 200$ models or more. Figure 12 (lower right) shows the confusion matrix for $N = 200$ and a feature length of 768 seconds.

		Classification (k-means)				
		preparing for work	going shopping	doing housework	Sum	Recall
GT	preparing for work	**7652**	921	916	9489	**80.6%**
	going shopping	1030	**6683**	1310	9023	**74.1%**
	doing housework	741	263	**14764**	15768	**93.6%**
	Sum	9423	7867	16990	34280	
	Precision	**81.2%**	**84.9%**	**86.9%**		

		Classification (Occ. Stats + NN)				
		preparing for work	going shopping	doing housework	Sum	Recall
GT	preparing for work	**1568**	336	72	1976	**79.4%**
	going shopping	86	**1669**	101	1856	**89.9%**
	doing housework	354	224	**2651**	3229	**82.1%**
	Sum	2008	2229	2824	7061	
	Precision	**78.1%**	**74.9%**	**93.9%**		

		Classification (Occ. Stats + SVM)			Sum	Recall
GT	preparing for work	**1383**	132	0	1515	**91.3%**
	going shopping	62	**1359**	126	1547	**87.8%**
	doing housework	14	143	**2612**	2769	**94.3%**
	Sum	1459	1634	2738	5831	
	Precision	**94.8%**	**83.2%**	**95.4%**		

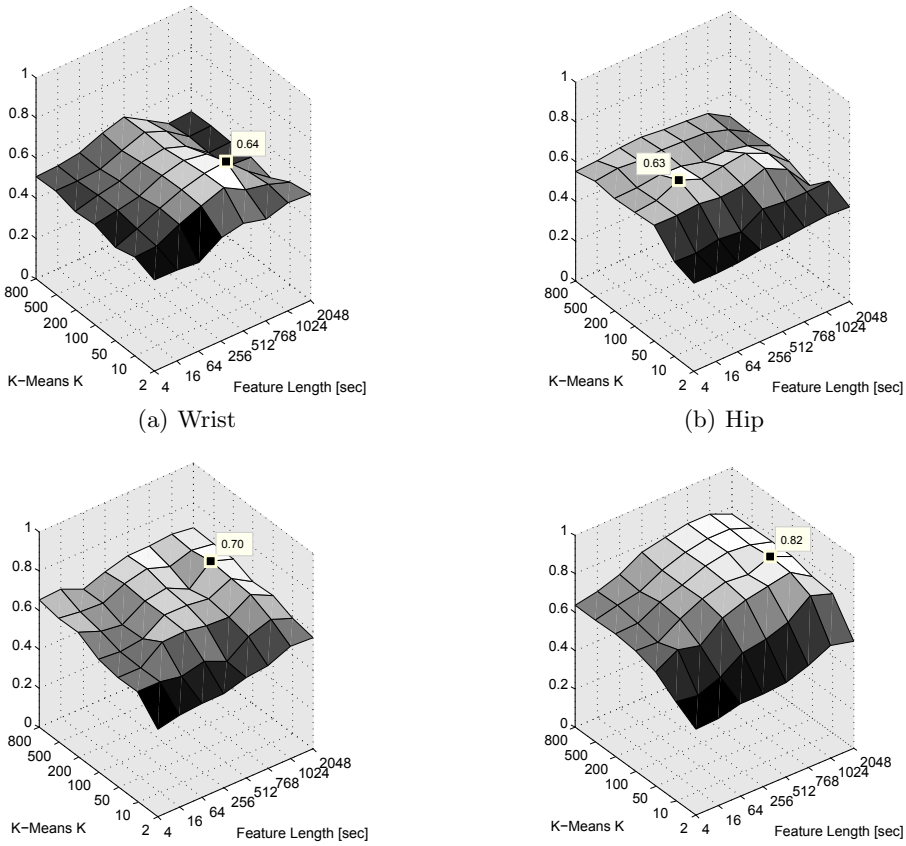
		Classification (HMM)			Sum	Recall
GT	preparing for work	**8220**	753	515	9488	**86.6%**
	going shopping	1042	**4962**	930	6934	**71.6%**
	doing housework	1156	536	**7334**	9026	**81.3%**
	Sum	10418	6251	8779	25448	
	Precision	**78.9%**	**79.4%**	**83.5%**		

Fig. 12. Aggregate confusion matrices for the best parameter combinations of the four approaches for recognizing high-level activities.

Sensor Placement. We also investigated the influence of different sensor locations on the recognition of high-level activities. Figure 13 shows the differences in performance when applying the k-means approach to subsets of sensors. Figure 13(a) shows the results for the wrist sensor. One can observe that for this sensor, the size of the feature window strongly influences the recognition rate – there is a distinct peak for relatively large windows between 512 and 1024 seconds. Obviously, for shorter windows the wrist movements are not discriminative enough for recognition. This might be due to the fact that the three scenes share some of the low-level activities, and that of these, many involve similar wrist movements, as for example *brushing teeth* or *showering*. The results for hip (Figure 13(b)) and thigh (Figure 13(c)) sensor do not exhibit such a clear tendency towards specific window lengths. Thus it appears that it is mainly the wrist sensor that is responsible for the good performance of relatively long windows when using all three sensors. The result for the hip sensor indicates that the performance at this location is more influenced by the number of clusters than the feature length; the best results are obtained for $k = 100$. Similarly as for the low-level activities, the combination of wrist and thigh sensor also performs very well for high level activities. For $k = 100$ and a feature length of 1024, the accuracy is 82%, i.e. only 3% worse than when using all three sensors.

6.1 Discussion

Figure 14 shows a summary table comparing the best results of the four approaches. As for the low-level activities, one observes that the approach *Occurrence Statistics + SVM* achieves the highest accuracy, in this case 91.8%. Combining the histogram features with an SVM instead of a nearest neighbor classifier leads to higher precision and recall values for all activities. Generally, the accuracy of all four approaches is over 80%, which is significantly higher than the chance level of about 33%. Even though the results might not generalize due to the small number of high-level activities in our set, we find that the high recognition rates are remarkable, considering the use of simple and easy-to-compute features in combination with a relatively large and challenging dataset.



(a) Wrist (b) Hip

(c) Thigh (d) Wrist + Thigh

Fig. 13. K-means based recognition accuracy of high-level activities for subsets of sensor locations. The best values of each combination are highlighted.

7 Conclusion

The main goal of this paper was to investigate how well current approaches in activity recognition can be applied to the recognition of high-level activities, which happen on the order of hours rather than minutes and consist of a diverse set of small scale activities. To this end, we recorded a naturalistic dataset with a user wearing three sensors on wrist, hip and thigh performing several instances of three different high-level scenes. We evaluated four different algorithms with respect to their ability to recognize both the low- and high-level activities contained in the dataset. One important aim of our research is to investigate to which extent current approaches for recognition of low-level activities can be directly applied to the recognition of high-level activities – i.e. using the same simple features without adding any intermediate levels of representation. We

	K-means		Occ./NN		Occ./SVM		HMM	
Scene	p	r	p	r	p	r	p	r
Preparing for Work	81.2	80.6	78.1	79.4	**94.8**	**91.3**	78.9	86.6
Going Shopping	84.9	74.1	74.9	89.9	**83.2**	**87.8**	79.4	71.6
Doing Housework	86.9	93.6	93.9	82.1	**95.4**	**94.3**	83.5	81.3
Mean	84.4	82.8	82.3	83.8	**91.1**	**91.2**	80.6	79.8
Accuracy	84.9		83.4		**91.8**		80.6	

Fig. 14. Summary of the results for high-level activities. The columns show the precision (p) and recall (r) values for each activity, as well as the accuracy.

believe that in the future such an approach would allow for scalable and efficient activity recognition systems based on simple sensors.

The results indicate that our algorithms can achieve competitive recognition rates for many of the low-level activities. The best results of slightly below 80% were achieved when using histograms of cluster assignments as features, combined with a support vector machine for classification. We investigated different window lengths and numbers of clusters and found that mapping the data to 50 clusters already leads to good results. In terms of sensor placement, using only two sensors at wrist and thigh resulted in equal or even better rates than using all three sensors.

When classifiying high-level activites, we achieve recognition rates of up to 92%, which is clearly above the chance level of about 33%. We achieve these results with the same algorithms that we used for the low-level activities, merely by changing parameters such as the feature length and classification window. The best results were again obtained by the histogram-based approach in combination with an SVM. For all our approaches we use simple mean and variance features derived from accelerometer readings at 2 Hz. Considering the relatively simple sensors and features, as well as the challenging dataset, we find that the results for the high-level activities are surprisingly good.

We conclude that recognizing activities on such scales using only small and unobtrusive body-worn accelerometers is a viable path worth pursuing. Yet we are aware that our work is but a first step towards recognition of high-level activities, and that more sophisticated models might yield better results. An obvious extension would be an hierarchical approach, using the outcome of the low-level classification as basis for the high-level inference, e.g. by defining a grammar of low-level activities. High-level activities however are often unstructured and may contain seemingly unrelated low-level activities, as e.g. observed in the data collection of this paper (e.g. when the user decided to eat during his housework). Therefore such an hierarchical approach is beyond the scope of this paper and will be explored in future work. In addition, we intend to validate our results on larger and more diverse sets of high-level activities, as well as across different users, in order to find out how well our approach generalizes.

Acknowledgements. This work is supported by the European Commission funded project MOBVIS (FP6-511051). Ulf Blanke gratefully acknowledges the scholarship provided by the DFG research training group "Topology of Technology".

References

1. Lester, J., Choudhury, T., Borriello, G.: A practical approach to recognizing physical activities. In: Proc. Pervasive. (2006)
2. Minnen, D., Starner, T., Essa, I., Isbell, C.: Discovering characteristic actions from on-body sensor data. In: Proc. ISWC. (2006)
3. Wyatt, D., Philipose, M., Choudhury, T.: Unsupervised Activity Recognition Using Automatically Mined Common Sense. Proc. AAAI 2005 (2005)
4. Huynh, T., Schiele, B.: Unsupervised discovery of structure in activity data using multiple eigenspaces. In: Proc. LoCA, Dublin, Ireland (2006)
5. Stiefmeier, T., Ogris, G., Junker, H., Lukowicz, P., Tröster, G.: Combining motion sensors and ultrasonic hands tracking for continuous activity recognition in a maintenance scenario. In: Proc. ISWC. (2006)
6. Wang, S., Pentney, W., Popescu, A., Choudhury, T., Philipose, M.: Common Sense Based Joint Training of Human Activity Recognizers. In: Proc. IJCAI. (2007)
7. Laerhoven, K.V., Gellersen, H.W.: Spine versus porcupine: A study in distributed wearable activity recognition. In: Proc. ISWC, Washington DC, USA (2004)
8. Bao, L., Intille, S.: Activity recognition from user-annotated acceleration data. In: Proc. Pervasive, Vienna, Austria (2004) 1–17
9. Patterson, D., Fox, D., Kautz, H., Philipose, M.: Fine-grained activity recognition by aggregating abstract object usage. In: Proc. ISWC. (2005) 44–51
10. Clarkson, B., Pentland, A.: Unsupervised clustering of ambulatory audio and video. In: icassp. (1999)
11. Eagle, N., Pentland, A.: Reality mining: sensing complex social systems. Personal and Ubiquitous Computing **10**(4) (2006) 255–268
12. Liao, L., Kautz, H., Fox, D.: Learning and inferring Transportation Routines. In: Proc. AAAI. (2004)
13. Krumm, J., Horvitz, E.: Predestination: Inferring Destinations from Partial Trajectories. In: Proc. UbiComp. (2006)
14. Marmasse, N., Schmandt, C.: Location-Aware Information Delivery with ComMotion. Proceedings of HUC 2000 (2000) 157–171
15. Van Laerhoven, K., Gellersen, H., Malliaris, Y.: Long-Term Activity Monitoring with a Wearable Sensor Node. Body Sensor Networks Workshop (2006)
16. Lo, B., Thiemjarus, S., King, R., Yang, G.: Body Sensor Network–A Wireless Sensor Platform for Pervasive Healthcare Monitoring. In: Proc. Pervasive. (2005)
17. Kern, N.: Multi-Sensor Context-Awareness for Wearable Computing. PhD thesis, TU Darmstadt (2005)
18. Ravi, N., Dandekar, N., Mysore, P., Littman, M.: Activity recognition from accelerometer data. Proc. IAAI (2005)
19. Huynh, T., Schiele, B.: Towards less supervision in activity recognition from wearable sensors. In: Proc. ISWC, Montreux, Switzerland (2006)
20. Oliver, N., Horvitz, E., Garg, A.: Layered representations for human activity recognition. Proc. ICMI (2002)
21. Lester, J., Choudhury, T., Kern, N., Borriello, G., Hannford, B.: A hybrid discriminative/generative approach for modeling human activities. In: Proc. IJCAI, Edinburgh, United Kingdom (2005) 766–772