Fine-Grained Activity Recognition of Pedestrians Travelling by Subway

Marco Maier and Florian Dorfmeister

Mobile and Distributed Systems Group
Ludwig-Maximilians-University Munich
Oettingenstraße 67, 80538 Munich, Germany
{marco.maier,florian.dorfmeister}@ifi.lmu.de

Abstract. With the now widespread usage of increasingly powerful smartphones, pro-active, context-aware, and thereby unobstrusive applications have become possible. A user's current activity is a primary piece of contextual information, and especially in urban areas, a user's current mode of transport is an important part of her activity. A lot of research has been conducted on automatically recognizing different means of transport, but up to know, no attempt has been made to perform a fine-grained classification of different activities related to travelling by local public transport.

In this work, we present an approach to recognize 17 different activities related to travelling by subway. We use only the sensor technology available in modern mobile phones and achieve a high classification accuracy of over 90%, without requiring a specific carrying position of the device. We discuss the usefulness of different sensors and computed features, and identify individual characteristics of the considered activities.

Keywords: Mode of Transport Recognition, Mobile Phone, Context Awareness, Activity Recognition.

1 Introduction

Context-aware applications and services have the ability to adapt to a user's environment. Although the term has already been coined in 1994 by Schilit and Theimer [6], only since the advent of smart mobile devices like today's smartphones, we can see a more wide-spread move towards context-aware computing. Besides *location*, *identity* and *time*, *activity* is regarded as a primary piece of information for characterizing a user's context [1].

There are many possibilities to get to know a user's current activity, ranging from manual input by the user to automatic recognition. The latter has been a major research topic for several years, but it many ways still is limited to very basic activities.

The first wave of attempts to perform automatic activity recognition often required the attachment of dedicated sensors like accelerometers or heart rate monitors [10] to the human body. However, with the proliferation of smartphones with their integrated sensors such as gyroscopes, compasses or barometers, it has

G. Memmi and U. Blanke (Eds.): MobiCASE 2013, LNICST 130, pp. 122–139, 2014.

become feasible to recognize a user's activity without any additional devices, promising a more convenient and ubiquitous user experience.

1.1 Mode of Transport Recognition

Especially in urban areas, a key information of a user's activity is her current *mode of transportation*, e.g., whether she is walking, cycling, driving a car, or travelling by any means of public transport like bus or subway. One can imagine a multitude of use cases including personal activity and health monitoring (e.g., *quantified self-tracking* [9]), creating ecological profiles of one's travelling habits, and first and foremost applications and services which automatically adapt their functionality to the user's current mode of transportation (e.g., interactive maps which focus on subway lines when travelling by subway).

There have been several attempts to automatically distinguish between different modes of transportation, which perform reasonably well. However, to the best of our knowledge, to date there exists no work aimed at recognizing activities and transportation phases on a finer-grained level, e.g., to tell apart entering, being on and exiting a subway train.

1.2 Fine-Grained Activity Recognition When Using the Subway

In this work we focus on recognizing several fine-grained activities and transportation phases when travelling by subway, i.e., walking in the subway station, walking upstairs/downstairs, using an escalator (up and down without walking, up and down while walking), using an elevator (up and down), waiting, waiting while the subway arrives, entering the subway train, standing in the subway while parking/accelerating/driving/decelerating, and exiting the subway train. In total, we try to recognize 17 different activities.

The subway as a means of transportation is interesting not only because it is an essential part of public transport in larger cities, but also because it is subject to restrictions such as the unavailability of GPS-based positioning. Beneath the previously mentioned use cases, finer-grained activity recognition while travelling by subway could enable pro-active services like reminding of getting off the subway train, or offering paperless ticketing on public transport (using recognized activities either directly for tracking/billing or indirectly as a means of fraud detection).

Being able to recognize activities and transportation phases while travelling by subway not only enables activity-aware services but also might solve the inherent positioning problem without GPS below ground: Patterns of entering a subway, accelerating, driving, decelerating and exiting could be matched to maps and subway timetables, leading to estimates about the user's current position.

1.3 Preconditions, Requirements and Contributions

We state the following preconditions and requirements for this attempt to fine-grained activity and transport phase recognition:

1. The solution should be realizable with the sensor technology of current smartphones, without any additional peripheral devices.
2. The user should not be required to carry the smartphone in a specific position.
3. Recognition of the current activity should be completed in a short timeframe, i.e., analyzing a large interval of sensor data afterwards in an offline manner is *not* sufficient[1].

In the following, we present a solution for this problem statement. By employing a supervised machine learning approach, we recognize 17 different activities with a high accuracy of over 90%. In our concept and evaluation, we try to infer qualitative statements which are generalizable to other scenarios and use cases. We include experimental results, and we describe and explain our experiences and observations regarding

- how sensor data should be collected and pre-processed
- which types of sensor data correlate with which activities in what way
- which computated features are suitable to represent these correlations

The rest of this paper is structured as follows: In section 2, our experimental setup and the data collection process is described. After that, we outline the general concept of our approach (section 3). In section 4, we describe insights regarding individual activities and their correlation with specific sensor data and features, and we assess the performance of our recognition approach. In section 5, we have a look at related work in the field of activity and transportation mode recognition, before we finish with a conclusion and an outlook at future work (section 6).

2 Experimental Setup and Data Set

In order to examine the characteristics of different activities and transportation phases of the subway, we collected a data set in the subway system of Munich, Germany. In this section, we explain the data collection procedure as well as the properties of the resulting data sets.

2.1 Hardware and Logging Application

We used Android-based Google Nexus 4 smartphones as test devices to collect the sensor data. Up to four devices were used in parallel to capture sensor data at different positions at the human body.[2] Each device was running a custom made logging application, recording the following information and sensor data:

[1] This requirement does not rule out offline training phases of machine learning algorithms. However, our goal is that once trained, the solution should not require more sensor data than what is available in a reasonable timeframe for near-realtime usage.

[2] Notice that we collected this data at four different positions only for the purpose of performance comparison, not for sensor fusion algorithms or the like.

- timestamp
- accelerometer (three axis)
- gyroscope (three axis)
- magnetometer (three axis)
- barometer
- GPS
- audio (microphone)

The sampling frequency was set to the maximum value (i.e., the corresponding listener of the Android application was set to `SENSOR_DELAY_FASTEST`).

To synchronize the recordings of all devices, they were linked via WLAN. We used an additional smartphone as the *master* device which set up the WLAN and was used to broadcast the current activity label[3]. The latter was done by sending a message via UDP to the *client* devices when a new label was selected at the master device. Client devices were required to confirm the new label, in order to ensure a data collection and labeling process as accurate as possible.

2.2 Data Collection

As stated in section 1.3, we aim for a solution which is independent of the carrying position of the smartphone. Therefore, we equipped a male test person with four devices, carried at the following positions:

- left front shirt pocket
- right front trouser pocket
- lying in the backpack
- held in the right hand

Additionally, the test person was holding the master device in his left hand, which he used to assign a label to the current situation. The test person could quite easily perform the labeling without bigger distractions, both mentally and concerning the other smartphones' sensor data readings.

2.3 Activities

We used a set of 17 different activities which we regard as important in the given scenario. The corresponding labels are summed up in table 1. The activities include those which can happen anytime outside the subway train (i.e., walking, waiting), typically occur when entering or leaving the subway station (i.e., walking downstairs or upstairs, maybe using escalators or elevators), and those which are linked to the transportation phases of the subway (i.e., entering, accelerating, driving, decelerating, exiting).

[3] The activity labels were used for our supervised learning approach later on.

Table 1. Overview of labels/activities which were recorded

label	meaning
walk	walking outside the subway, on even ground
walkup	walking upstairs
walkdown	walking downstairs
rollupwalk	walking on escalator, upwards
rolldownwalk	walking on escalator, downwards
rollup	standing on escalator driving upwards
rolldown	standing on escalator driving downwards
walkwait	standing still, on static ground
driveup	using elevator driving upwards
drivedown	using elevator driving downwards
subarrive	standing still (waiting) while subway train arrives
subenter	walking into the subway train
subwait	standing in the subway while not driving
subaccel	standing in the subway while accelerating
subdrive	standing in the subway while driving
subbrake	standing in the subway decelerating
subexit	walking out of the subway

2.4 Final Datasets

We performed three test drives in Munich. Experiments were started above ground, then entering the subway station, getting on an subway train, driving an arbitrary number of stations, exiting the train, returning above ground, re-entering the subway station, and so on.

In total, we collected 279 minutes of travelling by subway in Munich (in which the label changed 717 times). The pre-dominant label (*walk*) comprises 32.9 % of the collected dataset, i.e., simply always guessing that label would result in such a classification accuracy (which therefore serves as a baseline to be beaten by a more sophisticated approach).

3 Concept

We use a supervised machine learning approach to recognize the activities. After a suitable data set has been collected, such an approach typically requires the following steps:

1. Pre-Processing
2. Windowing
3. Cleansing
4. Feature Computation
5. Feature Selection
6. Choosing a Classifier

In the following, we describe our methods and choices in each of these steps.

3.1 Pre-processing

The hardware sensors on current Android devices typically have a sampling rate of about 20 to 80 Hz. Combined with the uncertainty of when different sensors return a new measurement, the actual sampling rate is fluctuating. In order to get a more stable sampling rate, we transformed the recorded values into an equidistant series of measurements. By filling in the missing values, we arrived at a synthetic sampling rate of 1000 Hz. It is important to note that we only performed this "up-sampling" to make feature computation easier. We did *not* employ any features using higher sampling frequencies than the sensors physically could provide.

Since the devices were carried at various orientations, the absolute values of the three axis of the accelerometer, the gyroscope and the magnetometer were not really meaningful. In each case, we combined the three values into a orientation-independent value by computing the norm of the three-element vector (i.e., $a = \sqrt{{a_x}^2 + {a_y}^2 + {a_z}^2}$).

3.2 Windowing

Raw measurements of sensor data usually cannot be used directly with machine learning algorithms but have to be combined by computing features on larger data intervals of certain length. In related work for transportation mode recognition, such windows often are eight seconds or longer [11]. Approaches for pure activity recognition typically use shorter windows like one or two seconds [13]. Regarding requirement 3 of section 1.3, windows should not be too wide, so that the lag for recognizing an activity does not get too long. However, they have to be big enough to capture periodicity in movements, etc. Therefore, we experimented with windows of length 1024 ms, 2048 ms and 4096 ms.[4] In each case, windows were overlapping 50%, which is a common approach in related work.

3.3 Cleansing

The windowing procedure sometimes resulted in windows with multiple labels assigned. These were removed from the dataset. Some recordings furthermore had unlabeled windows at their beginning (due to a lag between starting the recording and assigning the first label), which were also deleted.

3.4 Feature Computation

We computed a large set of features for each window to be able to examine the correlations of certain activities with certain features. There are features in both the time domain and the frequency domain (after performing an FFT on the respective data). We considered the following components:

[4] Using a multiple of 2 increases computation performance of the FFT used later on.

- norm of acceleration values
- norm of gyroscope values
- norm of magnetometer values
- pressure

and computed the following statistical features for each of them

- maximum, minimum, mean, standard deviation, 75th percentile
- root mean square
- number of zero crossings

For each combination, we also computed the difference to the respective value of each of the previous ten windows and of the following window. We then performed an FFT on each of the four components as well as on the audio data, and computed the following measures on the obtained coefficients:

- maximum, minimum, mean, standard deviation, 75th percentile
- dominant frequency
- root mean square
- energy, defined as
$$\frac{1}{n}\sum_{i=1}^{n} {x_i}^2$$
- entropy, defined as
$$-\sum_{i=1}^{n} P(x_i) log_b P(x_i)$$
- ratio of the mean to the mean of the complete spectrum
- ratio of the energy to the energy of the complete spectrum
- ratio of the entropy to the entropy of the complete spectrum

The above values were not only computed on the respective complete frequency spectrum but on the frequency intervals

1-3, 3-5, 5-8, 8-11, 11-16, 16-22, 22-29, 29-37 and 37-50 Hz

for the sensor data, and on the frequency intervals

1-20, 20-50, 50-100, 100-200, 200-500, 500-900, 900-1400, 1400-2000, 2000-2700, 2700-4000, 1-500, 500-1500 and 1500-4000 Hz

for the audio data.

The frequency intervals for the audio features partly were chosen according to certain characteristics we observed while visually investigating the recorded audio data. Figure 1 shows a short section of one of the audio recordings, visualized as a spectrogram.

There a several interesting moments in that recording. At time (1), we marked the beginning of a repeated signal tone (*beep*) which is played before the subway train's doors close. These short tones result in the signal peaks at about

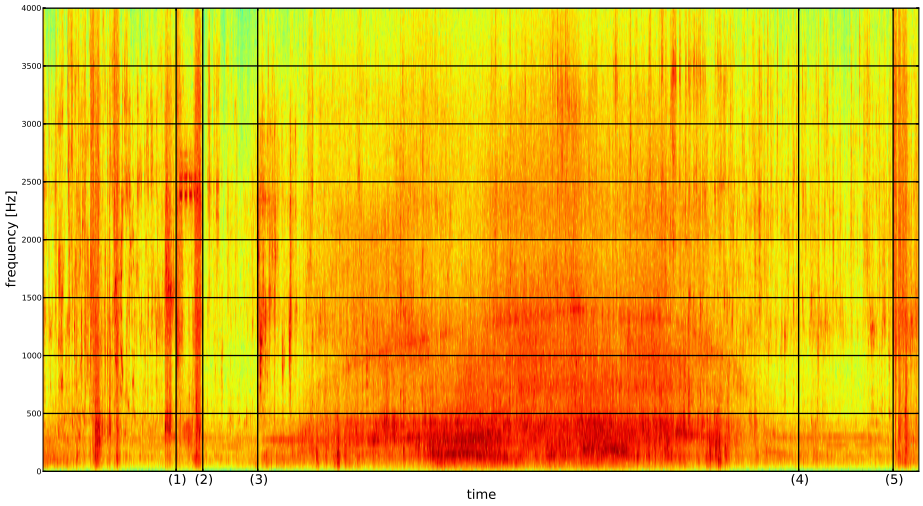


Fig. 1. Spectrogram of the audio signal recorded while driving by subway. At (1), the signal tone of the subway train is played (indicating doors closing). At (2), doors are closed. From (3) on, the subway train is accelerating, then driving and then decelerating until it stops at about (4). At (5), doors are opened.

2.5 kHz. After the signal, the doors close, indicated by a short "noisy" moment at time (2). This noise is caused by the pneumatic closing of the doors. Then there is a more silent period when the doors are closed but the subway train is still not moving. At time (3), the subway starts to accelerate, which results in an overall increased amount of energy in the signal spectrum. Interestingly, one can even visually conceive the sound of accelerating, driving and decelerating, which constitutes the near-semicircle in the frequency range from 1.0 kHz to 1.5 kHz. At about time (4), the subway has reached its parking position, again followed by some seconds of "silence" until the doors open. At time (5) the noise burst signifies the pneumatic opening of the doors.

Of course, these characteristics like the presence or the frequency of the signal tone can vary among traffic systems in different cities. However, the general idea to leverage these audio signatures can be adopted. Therefore, we not only included the aforementioned audio features but added some more features specifically targeted at the signal tone. To be more precise, we computed the maximum, mean, energy and root mean square of the frequency spectrum in the range of 2.4 kHz to 2.6kHz of the previous n windows ($n \in \{5, 10, 20, 30\}$) and added that values as features to the current window. The meaning of these features simply is that the "signal tone occured in the near past of the current window".

Preliminary tests showed that orientation-independet values of the gyroscope and the magnetometer do not provide any information beyond what is already observable in the accelerometer data. Therefore, we did not include those components in our evaluation. Making use of the gyroscope and magnetometer might be interesting when the position of the smartphone is known or is inferred in a pre-processing step [2].

3.5 Feature Selection

In total, we started with a set of 632 features, based on accelerometer, barometer and audio data. This large number of features on only a few raw data sources naturally tends to providing redundant information. Therefore, we reduced the number of features by performing a correlation-based feature subset selection [3] which resulted in a set of about 80 features which were used for the following evaluation.

3.6 Choosing a Classifier

We experimented with several types of classification algorithms, namely decision trees (J48), support-vector machines (SMO), bayesian networks, instance-based learning (kNN) and random forests. Preliminary tests showed best results with the random forest classifier, which therefore was used for most of the evaluation (where not stated otherwise).

4 Evaluation

In this section, we explain the influence of the window size as well as of the carrying position of the smartphone. After that, the merit of including the audio features is examined. We outline the usefulness of a hierarchical classification approach, show general results of different classifiers, and finally have a look at some interesting features which led to a good classification performance.

The evaluation was performed using the WEKA data mining software [4], and all the results have been obtained doing 10-fold cross validations on our dataset.

4.1 Window Size

We evaluated the performance of the three window sizes 1024 ms, 2048 ms and 4096 ms. The results are shown in figure 2a. One can see a slight increase of correct classifications with growing window sizes. Regarding requirement 3 from section 1.3, we opted for a window size of 2048 ms as a compromise, since larger windows might not be useful in practice.

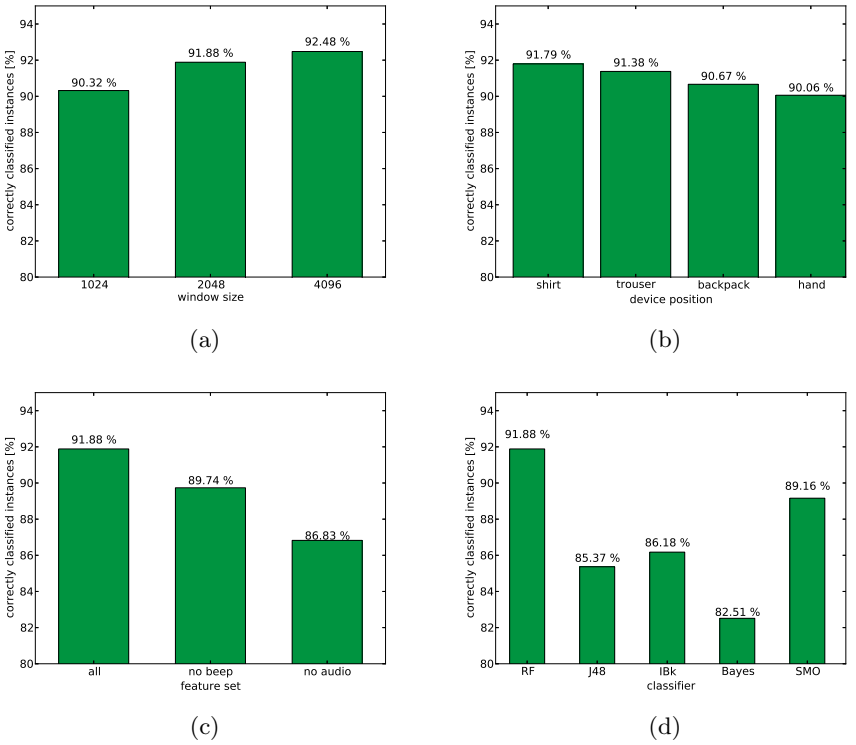


Fig. 2. Correctly classified instances by (a) window size, (b) device position, (c) feature set (with or without specific audio features) and (d) classifier (random forest (RF), decision tree (J48), k-nearest neighbour (IBk), bayes net (Bayes) and support-vector machine (SMO))

4.2 Carrying Positions

Regarding the different carrying positions, there is not much difference between the four choices (see figure 2b). The shirt pocket seems to be the most suitable position, probably due to a good combination of little mobility and less damping of the audio signal, e.g., compared to the trouser pocket. In general, a more stable position is better than a free one because this allows to better reflect the movement of the whole body.

Summing up, the chosen features with a special focus on orientation-independency allow for usage of the system in any of the tested positions.

4.3 Influence of Audio Features

As explained in section 3.5, some of the audio features were chosen based on observations in the frequency spectrum of the audio data. We therefore evaluated

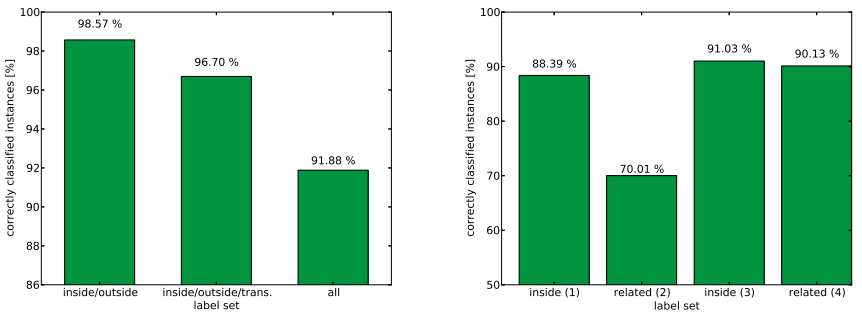
the influence of the audio features. Figure 2c shows the classification results when using

1. all features including audio
2. all features but without the *beep* features (i.e. those targeted at the signal tone)
3. all features without all the audio features

One can see a slight decrease in classification performance when removing the beep features. However, it seems that the general audio features can compensate this effect. Removing all the audio features leads to considerably inferior performance, proving that the audio signatures of the activities really are an essential part for good recognition results.

4.4 Classifiers

As explained before, we mostly used the random forest classifier in our experiments. Figure 2d shows the performance of four other classifiers compared to the random forest approach. All tested classifiers, namely decision tree, k-nearest neighbour, bayesian network and support-vector machine yield inferior results than random forest. Note that we did not investigate different parameter settings very thouroughly, since the basic performance of the random forest algorithm was sufficient for our evaluations, which is concordant with related work. Thus, the other algorithms might be tweakable to produce results just as good as or even better than the random forest.



(a) (b)

Fig. 3. Correctly classified instances when (a) grouping instances together and (b) performing a hierarchical classification (3+4) compared to a flat classification (1+2)

4.5 Hierarchical Classification

Although the general recognition results look good with the number of correctly classified instances above 90%, the results are a bit skewed due to an uneven distribution of the observed activities. Naturally, the activity "walking" occurs far more often than e.g. entering or exiting a subway train.

In order to more thoroughly assess the recognition rates of certain activities, we define four subgroups of activities:

- *outside*, which are activities happening outside the subway train (i.e., *walk*, *walkup*, *walkdown*, *rollupwalk*, *rolldownwalk*, *walkwait*, *rollup*, *rolldown*, *driveup*, *drivedown*).
- *inside*, which are activities happening inside the subway train (i.e., *subwait*, *subaccel*, *subdrive*, *subbrake*).
- *transition*, which are activities happening between *outside* and *inside* activities (i.e., *subarrive*, *subenter*, *subexit*).
- *related*, which are activities related to the subway train (i.e., the union set of *inside* and *transition*).

In a first step, we replaced the individual labels of the activities with the respective choice of either *outside*, *inside* or *transition*. Figure 3a shows the classification results when only trying to distinguish between *outside* and *inside* activities, and when categorizing into *outside*, *inside* and *transition* activities.

In both cases, over 95% of all instances where classified correctly. Considering the unavoidable inaccurracy in the data collection and labeling process, we regard this value as a reasonable maximum. Thus, classification into the three main categories is as good as can be.

We then examined the classification performance of the labels contained in the *inside* and *related* subsets when performing the classification on the whole dataset. The results can be seen in figure 3b in bars (1) and (2). The *inside* category yields quite good results, whereas the recognition rate of the *related* category drops down to about 70%. The reason is that activities such as entering or exiting the subway can easily be confused with ordinary walking. Furthermore, the number of available training instances is quite low in that category, so the differences compared to the *walk* activity are hard to grasp.

In order to solve this problem we opted for a hierarchical approach. By first applying a coarse-grained categorization into *outside*, *inside* and *transition* activities, we can quite accurately identify and rule out *outside* activities and apply a second classification to groups of *inside* or *related* activities. Using this procedure, we obtain recognition rates of over 90% for both the *inside* or the more general *related* activities (which include the problematic labels such as *subenter* and *subexit*). The results are shown in figure 3b in bars (3) and (4).

4.6 Interesting Features

In the following, three interesting features will be discussed exemplarily because they either had a huge influence on classification performance or were more or less suprising.

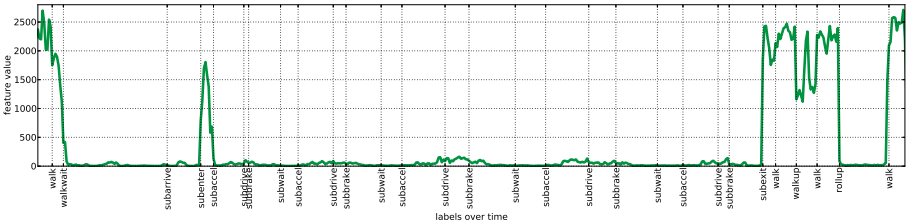


Fig. 4. Mean of frequency coefficients of accelerometer signal in range 1-3 Hz over time

In figure 4, one can see the mean value of the frequency coefficients of the accelerometer signal in the range from 1 to 3 Hz over time. This feature actually describes how present the frequencies from 1 to 3 Hz are within the window. Trying to group activities for which this feature exhibits larger values and those which only show values near zero, one can identify the disjoint groups of activities which involve "walking" and activities which do not. The reason is that frequencies of that range correspond to the typical step frequency of pedestrians.

This is an important feature to tell apart the activities *subenter* and *subexit* (which exhibit "walking" characteristics) from the other activities contained in the *related* subset (which do not).

One of the most difficult differentiations is between the activites *subaccel* and *subbrake* because they both simply are a kind of acceleration. Since we cannot rely on a specific orientation of the mobile device, we cannot tell apart positive (*subaccel*) from negative acceleration (*subbrake*).

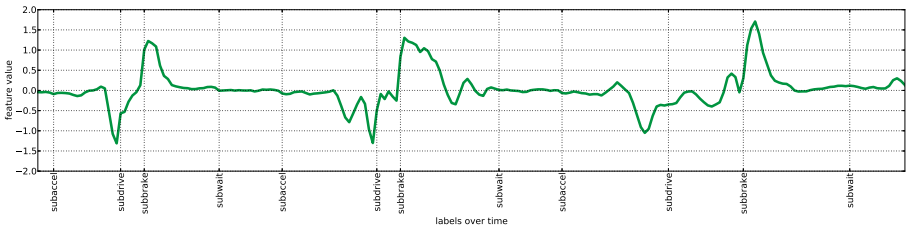


Fig. 5. Difference of mean pressure in window compared to the 3rd previous window over time

One feature to distinguish between these two activities is visualized in figure 5, which shows the difference of the mean pressure in the current window and the mean pressure in the 3rd previous window over time.[5]

[5] We observed the same pattern with each ith previous window, the 3rd previous one only showing the clearest peaks in this case.

In general, the features related to air pressure where included to better indentify activities which involve going up- or downstairs. However, we can also spot an interesting pattern concerning the activities *subaccel*, *subdrive* and *subbrake*. When accelerating, the pressure is going down (i.e., the difference to the previous window is negative). When driving, the ratio is normalizing to zero. Finally, when decelerating, the pressure is increasing (i.e., the difference to the previous window is positive).

This pattern is observable throughout our dataset. Nevertheless, we aim to investigate this behaviour more thouroughly in future work, since the pattern might be dependent on the exact location at which the user is standing within the subway (i.e., in the front or in the back).

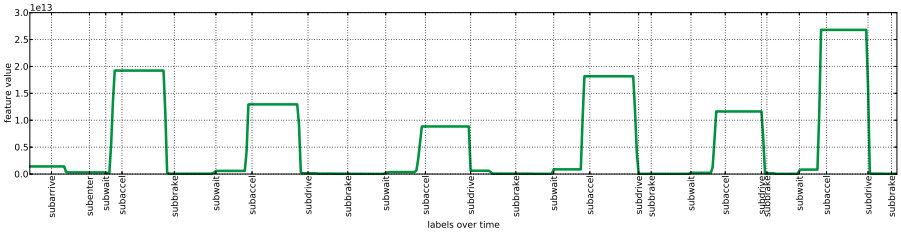


Fig. 6. Maximum energy of "beep" within last 30 windows over time

Another characteristic to tell apart the *subaccel*, *subdrive* and *subbrake* activities is the signal tone, as explained in section 3.4. Figure 6 shows the maximum energy of the signal tone's frequency range within a window's previous 30 windows over time. One can clearly see that the signal tone results in a peak in the following windows. Considering 30 windows of length 2048 ms with 50% overlap results in a timeframe of 30 seconds, which seems to correspond quite well to the typical acceleration length of the subway trains in Munich. Of course, this rather perfect mapping is biased, but the general idea is feasible.

In general, we found that several distinct features of both accelerometer and barometer values were chosen by the classifiers, with features concerning acceleration mostly being a variant of "energy in a certain frequency interval", and features concerning barometric pressure mostly being a variant of "difference to a certain previous window".

Summing up, we can state that our presented approach provides excellent results to recognize 17 important activities related to travelling by subway. Using position-independet features, we do not require a specific carrying position. A window size of 2048 ms enables responsive applications. We have shown that considering the audio signal leads to better results, and that a hierarchical classification approach renders high classification accuracy possible even for easily confusable activities such as "entering the subway train".

5 Related Work

In this section we have a look at existing approaches for activity recognition, especially the ones aiming at determining a user's current mode of transportation based on her mobile device's sensors.

Zhang et al. [13] aim at identifying the most important features for human activity classification, facing the problem that not all of the features available are equally useful for activity classification. Therefore, the authors evaluated the performance of three different algorithms for feature selection. The nine different activities the authors were trying to detect were walking forward, left, right, going upstairs and downstairs, as well as jumping, running, standing and sitting still. The recording device was attached to the users' hip, thereby somehow constraining the generality of the evaluation results by allowing for the assumption that the sensors location and orientation is known. In order to improve classification performance, the authors propose to use a multi-layer classification framework grouping activities into appropriate subsets and then performing feature selection and classification. This allows for using different features for different activity subsets, granting more flexibility, performance and accuracy. We adopted this idea for our finegrained activity classification (see section 4.5).

Yatani et al. [12] present *BodyScope*, which is a wearable acoustic sensor recording sounds from a user's throat area and classifying them into activities, such as eating, drinking, speaking, laughing, and coughing. Using a SVM-based classfication technique, the authors are reaching a combined F-score of about 79% for all of their twelve different activities. For classification, the authors use the zero-crossing-rate as a time-dependent feature, as well as several frequency-dependent features such as total spectrum power, brightness, spectral rolloff and spectral flux. With the F-measure resulting in 49.6% with the Leave-one-participant-out technique and 79.5% with Leave-one-sample-per-participant-out, the authors conclude that their classifier has to be trained individually on a per-user-basis.

Sun et al. [8] are making usage of the accelerometer readings in a current smartphone in order to monitor the daily activities of the smartphone's user. Just as we do, the authors put great effort in determining a user's current activity independent from the smartphone's orientation and different pocket locations. The features used are the three accelerometer axes as well as their magnitude's mean, variance, energy and entropy as well as the four components' correlation, summing up to 22 features per frame. The authors try to recognize seven everyday human activities such as being stationary, walking, running, bicycling, ascending and descending stairs as well as driving a car. Following a SVM based approach, the authors reach an overall F-score of over 93%. As their evaluation shows, however, classification results are even more accurate assuming that the pocket location is known in advance.

Quite similiar to our approach, Reddy et al. [5] aim at determining a user's current mode of transportation based on her mobile phone's sensor data readings alone. The authors are making usage of the phone's GPS module and accelerometer recordings in order to determine whether the user is sationary, walking,

running, biking or in motorized transport. As a means for classification, the authors deploy a decision tree, postprocessing its output with a first-order hidden markov model, resulting in classification accuracy over 90%. Just as with our approach, this work is not making any assumptions about the phone's pocket location and orientation, making it generically applicable. Features in use are the user's speed derived from her GPS module, as well as her accelerometer readings' energy, variance and the sum of FFT coefficients lower than 5 Hz. According to their evaluation, accelerometer and GPS data readings should be used complementary, leading to accuracy gains of up to 10%. With their two stage approach consisting of a decision tree and a HMM, the authors reach classification results over 98% accuracy. However, the authors are not able to differentiate between different kinds of motorized transport or making any fine-grained assumptions for any subactivities such as entering a vehicle or accelerating. Moreover, relying on GPS data such as the speed received from the GPS module, this approach is not applicable to underground activity classification.

Going one step further, Zhang et al. [14] try to make more fine-grained assumptions about a user's mode of transportation. They base their classification on both mobile phones and wearable foot force sensors. Hence, the authors are not only using a smartphone (GPS), but also some specialized kind of sensing hardware mounted to a user's feet. The different modalities they try to recognize are walking, cycling, as well as being a bus passenger, car driver and car passenger. Their activity classification is hence more detailled than Reddy's, but is also relying on the GPS module and a specialized sensor placed on the user's foot. They reach a 95% accuracy with 10 different individuals. In order to save on the smartphone's computation and battery capacities, the authors primarily focus on using time-domain features. These include mean, maximum and standard deviation of the GPS speed of a window, as well as mean, maximum and standard deviation of both feet's foot force sensors. The authors compared Naive Bayes, Decision tree and decision table techniques against each other. For all five modes of transportation, the decision tree allows for an overall classification accuracy of 97.3%. Evaluating the different motorized modes only, DT still reaches 87.5% accuracy.

Stenneth et al. [7] examine the possibility of transport mode detection using mobile phones and GIS data. The classification algorithm takes a smartphone's GPS readings and map information of the underlying transportation network as input. Based on these data, the authors try to determine different modes of transportation, namely by car, bus, train, walking, cycling and being stationary. In contrast to all other works, the authors additionaly provide real time information of time and location of public transport vehicles in order to achieve higher classification accuracy. As new and innovative classification features, the authors are making usage of average bus location closeness, candiate bus location closeness, average rail line trajectory closeness and bus stop closeness rate. Additionally, standard features such as average speed, average heading change, average acceleration and average accuracy of GPS coordinates are used. Using all available information in a Random Forest classifier, the authos are able to

reach an average classification accuracy of 93.7%. However, it is not possible to make fine-grained assumptions for a single mode of transportation or subway based transportation.

6 Conclusion and Future Work

In this work, we presented an approach to automatically recognize 17 important activities related to travelling by subway, using only the sensor technology of a modern smartphone. To the best of our knowledge, this is the first attempt to perform a really fine-grained recognition of different activities and phases of a user's mode of transport.

We achieve a high classification accuracy of over 90% even for problematic activities such as "entering the subway train", which can easily be confused with ordinary "walking", all without requiring a specific carrying position of the mobile device and while preserving the potential for near-realtime applications.

Despite the promising results, this work certainly is only a first step towards a deployable system. One major shortcoming is the lack of a larger data set comprising a lot more test users. Although we tried to focus on features which should be rather user-independent in theory, only a more thorough evaluation will prove that sentiment.

Another aspect which will be tackled in future work are the temporal dependencies amongst the individual activities. So far, we do not make use of facts like "entering the subway happens before exiting the subway". We are convinced that introducing a higher level, reasoning layer could be able to compensate for a drop of classification accuracy which might be observed when having a larger user base without that sophisticated training data as we had in our evaluation.

Finally, related aspects such as energy consumption, privacy and security of the system have to be considered in future work.

All in all, the presented work and our evaluation of the activities' characteristics with regard to certain features can be used as a basis for investigating fine-grained activity recognition in other fields, especially regarding other means of transportation.

Acknowledgement. We would like to thank our student Uwe Müller for collecting large parts of our data set during his bachelor thesis.

References

1. Abowd, G.D., Dey, A.K., Brown, P.J., Davies, N., Smith, M., Steggles, P.: Towards a better understanding of context and context-awareness. In: Gellersen, H.-W. (ed.) HUC 1999. LNCS, vol. 1707, pp. 304–307. Springer, Heidelberg (1999)
2. Fujinami, K., Kouchi, S.: Recognizing a mobile phone's storing position as a context of a device and a user. In: Zheng, K., Li, M., Jiang, H. (eds.) MobiQuitous 2012. LNICST, vol. 120, pp. 76–88. Springer, Heidelberg (2013)

3. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. PhD thesis, University of Waikato, Hamilton, New Zealand (1998)
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD Explorations Newsletter 11(1), 10–18 (2009)
5. Reddy, S., Burke, J., Estrin, D., Hansen, M., Srivastava, M.: Determining transportation mode on mobile phones. In: Proceedings of the 2008 12th IEEE International Symposium on Wearable Computers, ISWC 2008, pp. 25–28. IEEE Computer Society, Washington, DC (2008)
6. Schilit, B.N., Theimer, M.M.: Disseminating active map information to mobile hosts. IEEE Network 8(5), 22–32 (1994)
7. Stenneth, L., Wolfson, O., Yu, P.S., Xu, B.: Transportation mode detection using mobile phones and gis information. In: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS 2011, pp. 54–63. ACM, New York (2011)
8. Sun, L., Zhang, D., Li, B., Guo, B., Li, S.: Activity recognition on an accelerometer embedded mobile phone with varying positions and orientations. In: Yu, Z., Liscano, R., Chen, G., Zhang, D., Zhou, X. (eds.) UIC 2010. LNCS, vol. 6406, pp. 548–562. Springer, Heidelberg (2010)
9. Swan, M.: Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking. International Journal of Environmental Research and Public Health 6(2), 492–525 (2009)
10. Tapia, E.M., Intille, S.S., Haskell, W., Larson, K., Wright, J., King, A., Friedman, R.: Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. In: 2007 11th IEEE International Symposium on Wearable Computers, pp. 37–40. IEEE (2007)
11. Wang, S., Chen, C., Ma, J.: Accelerometer based transportation mode recognition on mobile phones. In: 2010 Asia-Pacific Conference on Wearable Computing Systems (APWCS), pp. 44–46. IEEE (2010)
12. Yatani, K., Truong, K.N.: Bodyscope: a wearable acoustic sensor for activity recognition. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp 2012, pp. 341–350. ACM, New York (2012)
13. Zhang, M., Sawchuk, A.A.: A feature selection-based framework for human activity recognition using wearable multimodal sensors. In: Proceedings of the 6th International Conference on Body Area Networks, pp. 92–98. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2011)
14. Zhang, Z., Poslad, S.: Fine-grained transportation mode recognition using mobile phones and foot force sensors. In: Zheng, K., Li, M., Jiang, H. (eds.) MobiQuitous 2012. LNICST, vol. 120, pp. 103–114. Springer, Heidelberg (2013)