Assessing Activity Recognition Feedback in Long-term Psychology Trials

Manuel Dietrich
Topology of Technology
TU Darmstadt, Germany
dietrich@gugw.tu-darmstadt.de

Eugen Berlin
AGT International
Darmstadt, Germany
eberlin@agtinternational.com

Kristof van Laerhoven
Embedded Systems
University of Freiburg
kristof@ese.uni-freiburg.de

ABSTRACT

The physical activities we perform throughout our daily lives tell a great deal about our goals, routines, and behavior, and as such, have been known for a while to be a key indicator for psychiatric disorders. This paper focuses on the use of a wristwatch with integrated inertial sensors. The algorithms that deal with the data from these sensors can automatically detect the activities that the patient performed from characteristic motion patterns. Such a system can be deployed for several weeks continuously and can thus provide the consulting psychiatrist an insight in their patient's behavior and changes thereof. Since these algorithms will never be flawless, however, a remaining question is how we can support the psychiatrist in assigning confidence to these automatic detections. To this end, we present a study where visualizations at three levels from a detection algorithm are used as feedback, and examine which of these are the most helpful in conveying what activities the patient has performed. Results show that just visualizing the classifier's output performs the best, but that user's confidence in these automated predictions can be boosted significantly by visualizing earlier pre-processing steps.

CCS Concepts

•**Human-centered computing** → **Visualization;** *Ubiquitous and mobile computing;*

Author Keywords

Context-aware services; Activity recognition; Interaction design; Visualization methods

INTRODUCTION

Automated recognition of a person's activities has in the past decade often been suggested as an attractive system for monitoring patients that suffer from a wide range of disorders. By relying on observations from sensors in the environment or from body-worn sensors, performed activities can be inferred through a ubiquitous system. Activity recognition is motivated by establishing a more effective dialogue between user and computer, reducing cognitive load in pervasive computing scenarios, and delivering an improved service by proactively responding to given situations. Numerous applications have been suggested to benefit from this: Philipose et al. [27] for instance demonstrate how Activities of Daily Living (ADLs) can be detected, to estimate the quality of self-care for elderly users. Further application scenarios for activity recognition include detecting office activities [24], maintenance tasks performed by engineers [28] and specific sports activities [32, 13], finding appropriate advertising based on the user's physical activity [25] and eating and drinking activities [1]. Depending on the application, algorithms can go beyond recognition of activities and detect certain characteristics, such as the number of counts for selected gym workouts [5].

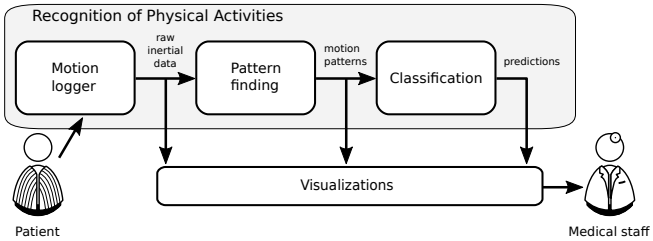

Figure 1. An activity recognition system (top) uses long-term inertial data from a patient's wrist-worn motion logger to find typical patterns matching that activity. These data are visualized and presented to consulting medical staff (right). This paper investigates what combinations of visualizations are the most helpful to convey past activities.

This paper's approach is targeting a class of applications that remains challenging: Among the more recent of activity recognition scenarios is psychiatric patient monitoring, which aims at following mood and behavioral trends by recording activity data over a period of typically *several months*. Existing commercial actigraphy solutions such as the MotionWatch [4] are able to record activity *levels* and to detect sleep and wake cycles for such long deployments, and come with tools for facilitating the recording of basic physical activity. In this scenario, some general problems in activity recognition are bypassed: Patients already keep *detailed diaries of their activities* so that supervised learning methods can be employed, and only *a few physical activities* linked to daily routine are of interest in the logged data. Other requirements, however, form novel challenges: Sensors should record for *long stretches of time*, a *large amount of logged data* needs to be analyzed, and detection needs to be *robust against a lot of background data*. This leads to recognition systems that will not always predict with 100% accuracy when an activity of interest happened. We

MUM '15, November 30-December 02, 2015, Linz, Austria

ISBN 978-1-4503-3605-5/15/11…$15.00

DOI: http://dx.doi.org/10.1145/2836041.2836052

argue that this can be reflected in the way the detections are visualized, by adding lower-level information visualizations.

Figure 1 illustrates the paper's healthcare scenario, based on collaborations with psychiatrists and psychologists guiding bipolar patients. Data are recorded per patient over several weeks and, after downloaded, can be visualized to the medical staff on three levels: (1) by plotting the raw data acquired by the wrist-worn inertial sensor (recording 3D accelerometer readings at 100Hz), (2) by displaying which sections of these data were identified as containing characteristic patterns for a particular activity, and (3) the final classifications made by the classification algorithm. We investigate which combination of these three information visualizations is most appropriate.

The remainder of this paper is structured as follows: First, a specific long-term monitoring scenario is described to motivate the need for a detection system that estimates when a user has performed certain physical activities. Then, the next section is dedicated to related work in activity recognition, with particular focus on methods that aim for long-term deployments in healthcare with feedback to others. The third section will present the details on our used method, and the levels of information abstraction at which the data can be visualized. In section 4 (Evaluation), the study is described, that uses a 33 participant, week-long *patient* dataset to visualize activities to a set of 15 *medical staff* independent participants, investigating the applicability of combinations of the three proposed visualizations for detected activities. The paper is wrapped up with the conclusions section enumerating the key results of this paper, as well as the future research potential.

CASE STUDY: BIPOLAR PATIENT MONITORING

We focus first and foremost on a practical capturing and detection system that is able to recognize particular activities within large stretches of time that tend to include a massive amount of motion and posture data, generally holding weeks of activity data at a time. As a case study of in which areas such a system would be applied, we start this section with a description of long-term monitoring of bipolar patients.

Research in mood disorders (such as attention deficit hyperactivity disorder (ADHD) and bipolar disorder [8]) relies frequently on the patients' self-reports, as well as semi-structured interviews with a psychiatrist, during diagnosis and therapy. Work with actigraphy tools in psychiatry [34, 30] has started to deploy wrist-worn sensors in conjunction with these tools that are recording the activity *intensities* observed for the patient from several seconds to minutes at a time.

Characterized by severe mood swings between manic or hypomanic, mixed, as well as depressive episodes, it is important in the diagnosis of a bipolar disorder to record the patient's activities over multiple weeks to months at a time. For mania for instance, energy levels tend to be high and activities tend to be performed in an interleaved fashion or especially vigorously. Similarly, depressions tend to correlate with lower activity levels, or in shortened stretches for physical activities, from not performing them at all or sparsely, to abandoning them. Apart from daily activities such as sleep and food intake, especially physical and leisure activities are very likely to be impacted: Patients might for instance decrease physical exercise during depressions, or vigorously practice playing a musical instrument for several hours in a manic episode.

The representation of performed activities over multiple weeks to a consulting psychiatrist is challenging not only because the activity recognition system cannot be guaranteed to work perfectly, but also because the person inspecting the results will have limited knowledge about the person who was wearing the sensor unit. Additionally, it is generally not feasible to rely on the self-recall of the patient over the course of several weeks for any corrections in the activity detections, since these can be expected to be biased.

As a precursor to our work, a series of interviews with psychiatrists resulted in a list of basic requirements that an activity recognition method should adhere to. These were grouped in three categories that are important to consider when designing a recognition system for psychiatric trials:

- **Supervised learning.** Patients are typically interviewed at regular intervals of several weeks, and provide log entries to report on performed tasks and their mood. Current actigraphs combine these reports with sensor data, so that the reports can be used to train patient-specific classifiers.

- **Week-long, 24/7 data.** Data needs to be captured at all hours of the day, as patients that go through depression or manic episodes are known to perform activities irregularly, including at night. The sensor units thus need to be robust and power-efficient to keep recording continuously, and the amount of data will be substantial to process.

- **Selected activities.** The number of activity classes to recognize is relatively small and can be determined by medical staff during the first set of interviews. This makes it easier for patients to keep a diary of which selected activities were performed, and means only few activities need to be detected amongst a large amount of background data that might produce false positives.

The next section will review literature on activity recognition methods that tackle similar approaches in wearable sensing and research on feedback about activity recognition results in a medical context.

RELATED WORK

Activity recognition has previously been suggested as a promising instrument for use in behavioral studies that capture events automatically beyond self-reports [23]. Both [33] and [29] have pointed out that the use of automatically monitored activities would be useful to support the diagnosis of bipolar disorder and detect onsets of depression and mania. In particular the so called Hamilton Depression Scale (HAMD) and Bech-Rafaelsen Mania scale (BRMS) [2] tools contain elements where physical activities are of considerable interest. To our knowledge no research has yet focused on an activity recording method that can be worn at the wrist for weeks and allows almost-instant analysis at the psychiatrist's office.

A significant amount of work in the context of activity recognition has focused on automatic feature selection for inertial

data and using strong classifiers upon these features to detect activities. Common candidates that have proven worthwhile in previous studies (e.g., [15], [19]) have found basic statistics, especially mean and variance, and frequency-based features (FFT and Cepstral Coefficients, spectral entropy and energy) over a sliding window to be distinctive features to characterize. Lester et al. [19] use in a combined discriminative-generative classification approach the AdaBoost algorithm to automatically select the best of these features and to learn an ensemble of static classifiers to recognize different activities. Strong classifiers that have proved valuable in activity recognition include Naïve Bayes, Bayesian Networks, Hidden Markov Models (HMMs) or Support Vector Machines (SVMs) [27, 24, 28, 13, 25, 1, 5, 19, 21, 26].

The use of motif discovery has been suggested as an alternative approach in activity recognition that is especially useful when a fully supervised method is not feasible, or when short characteristic gestures need to be spotted that are hard to annotate individually by the system's users. [22] use motifs to automatically discover gym work-out gestures in inertial data recorded form body-worn sensors, by mapping the sensor data to symbols and using a suffix tree to search efficiently through the resulting large symbolic strings. Similarly, [12] analyze activities in an instrumented kitchen, and [31] use motif discovery to detect activities such as walking and falling without supervision. We use motif discovery primarily because (1) the annotations that describe which activity was done when are provided by the system's wearer using self-recall and are thus only approximate, (2) we assume that a variety of physical activities can typically be characterized by occurrences of certain short gestures, and (3) because it is an especially fast method that allows parsing of large datasets at once.

Approaches on visualization methods have thus far played a less prominent role in activity recognition and are still underrepresented in the research. However, especially in a medical setting, in our case in which psychologists use activities as indicators for the behavior of patients, it is essential to have a valid presentation. A focus on the concepts of visualizations in the research can especially found in approaches where beside the facilitation of personal data also a further interest of motivation and persuasion is addressed. Examples can be found in "Fish'n'Steps", [20] a fish in a virtual tank which grows when doing fitness activity (step counter) as well as UbiFit Garden [7] which does the same with a flourishing garden. An alternative approach is done by [17] who visualize physical activity using 3D prints as habitual feedback. These projects can also be seen as gamification aspects of activity recognition. In most activity recognition publications, focus lies predominantly on improving the detection.

A systematic analysis and survey of existing personal data visualizations is done in [9] where different visualization heuristics are provided, especially for personal activity and behavior. As a suggestion for how visualization can support the understanding of the detected information, they use financial analytics. Other projects focus on alternative ways how to visualize physical activity data, for instance as a spiral view on the activities which enables a data survey over a long period of time [18]. As well [10] investigate different visualizations in self-tracking applications with the focus on comparing results and finding correlations. All these examples are approaches that highlight the role of visualization in activity recognition; Projects addressing health care applications in a broader sense are mostly about the motivation aspect which is realized with concepts of gamification. One project which has a similar constellation is these of [14] focusing on physical therapy by tracking activity data and visualizing it for a therapist.

Figure 3. The custom-built logging platform has form factor and function of a wristwatch, logging inertial motion data at 100Hz for weeks.

Another important aspect of our research is the long-term recording of inertial data, in an unobtrusive manner. This has been stressed in several key publications on activity recognition (most notably [6, 16]). Although datasets have been recorded over similar time frames as in this paper, none so far have recorded 24/7 for many days consecutively. Actigraphy has as an advantage that it does log for extended periods of time, but it abstracts the inertial data on the sensor unit and does not retain the original time series at a resolution that facilitates fine-grained activity modeling.

DETECTING ACTIVITIES WITH A WRIST-WORN LOGGER

This section gives an overview of the functionality of the detection system which is the base for the three types of visualization shown in Figure 1. The amount of raw data is reduced by a preprocessing step to make it easily presentable to the user and processable in the next steps. The second step, performing pattern recognition, is based on automatically detecting characteristic motifs. Motifs are typical substrings for an activity of interest, the preprocessed data is searched. In a last step, the actual classification into activities, the accumulated motifs are used to find out whether a certain activity has taken place at that time. We will present the technical background of the recognition process regarding these three steps. A detailed description of the method and a study regarding the detection accuracy can be found in [3].

Data Logging Platform

The hardware (see Figure 3) was designed to capture inertial data at a relatively fast rate of 100Hz for long-term deployments, where the user can wear the logger day and night. At the center of the platform is a low-power microcontroller (Microchip 18F46j50) that obtains 3D acceleration readings by an attached ADXL345 accelerometer, which are then suitably compressed and stored on a micro-SD card for later analysis. The unit can be connected via its mini-USB port to a host computer for data downloading and recharging its 180mAh miniature battery. The display can be activated by double-tapping to show the current time and date (though for our experiment other, display-less units were also used that tended to be more robust for everyday use).

Raw Inertial Data (mSWAB)

The first abstraction step is essential from an efficiency point of view: Since the accelerometer sensor is sampled at 100Hz for capturing the essence of the gestures and typical motions

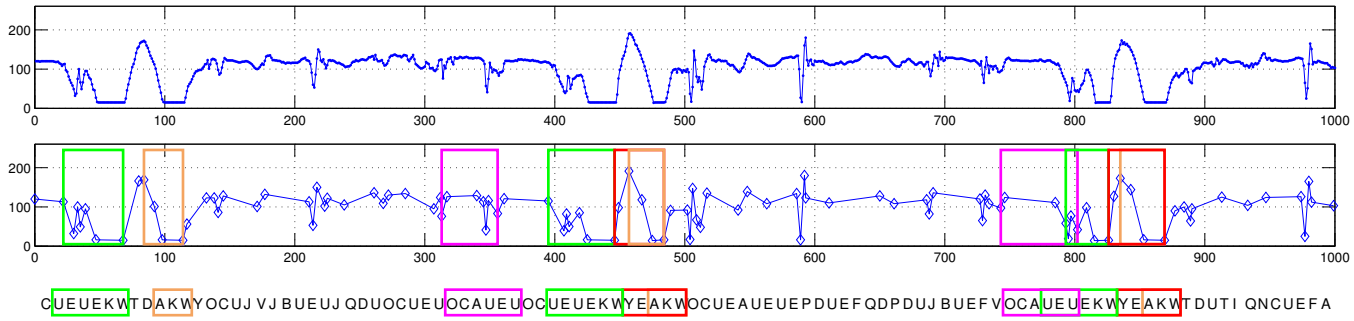

Figure 2. The raw inertial data (top plot) are transformed by a piecewise linear approximation algorithm into segments (bottom plot) that preserve the shape of the signal to facilitate storage and analysis. The segments are subsequently abstracted in discrete symbols to allow fast discovery and matching of motifs. Occurrences for four motifs are highlighted by colored boxes; Note that we allow overlaps and variations in length.

performed by the sensor's wearer, this also means that the dataset size grows quickly and becomes computationally challenging, both for analysis and for visualization. We argue that for discovering characteristic gestures within inertial activity data, primarily the shape of the signal is important to retain.

For our system we use a modification of the Sliding Window and Bottom-Up algorithm (mSWAB) that has been verified to perform well on body-worn accelerometer data [32]. The result of this step is a linear approximation of the raw inertial data that is typically an order of magnitude smaller than the raw data [3]. An example of such a transformation can be found in top part of Figure 2.

Finding recurring characteristic patterns (Motifs)

After the abstraction of the raw acceleration data to linear segments, the pattern recognition is realized by first mapping the segments to a symbolic representation and second detecting efficiently recurring substrings within this string of symbols as characteristic patterns (also called motifs in data mining literature). The symbolic representation and the merging motifs are illustrated in the second plot in Figure 2. The mapping of segments to a symbols uses subsequent angles between pairs of the linear segments to efficiently represent peaks in the data. Having mapped the raw acceleration data to a symbol sequence, an approach called motif discovery can now be used to find substrings that occur multiple times in the target class. This is above all an efficiency problem: searching for all occurrences of every motif in a long string in an exhaustive and brute-force fashion will result in a slow discovery process that is not scalable, as large sets of motifs are expected to be present. Instead, a data structure called suffix tree is used, which can be constructed in linear time and is able to represent all possible suffixes in the string as a tree: Determining motifs is thus reduced to traversing the tree from root to a certain depth, at which the suffix tree has stored all possible instances at which this substring is found.

Classification of Activities (Dense Motif Discovery)

Using the most discriminant motifs for a given activity class during a training phase, classification is performed by local evidence of all motifs that support an activity. This is implemented using a bag-of-words classifier over a sliding time window that traverses the time series and accumulates local evidence by straightforward counting of the motif occurrences. As the activities tend to last approximately 60 minutes, a window size of 10 minutes was chosen. The resulting classifier has shown to produce results in line with state-of-the-art activity recognition research with good predictions for many physical activities [3]; Important to note is also the fact that the execution of the above three-step process for classifying raw inertial data to activities takes maximally a few minutes on a standard computer, thus enabling an almost immediate visualization after the data is uploaded when the patient visits the consulting psychiatrist. A more detailed listing of detection measures can be found in Table 1.

Visualization of the Abstraction Steps

The previous three abstraction steps used by the system to convert raw inertial data to activity classes can be visualized on a horizontal time axis as already illustrated in Figure 2. We have opted to keep the visualizations as clean as possible without annotations to the axes, and have unified all plots to keep the x axis span to exactly 24 hours (see Figure 4). In the evaluation, we will refer to these single visualizations as *A*, *B*, and *C* respectively, and will note combined visualizations with the + operator (e.g., we refer to *A*+*C* when the raw inertial data and the classification were shown). In the evaluation section of this paper, we thus investigate which of these combinations helps the user the most to assess when an activity occurred.

EVALUATION

We are in this study especially interested in how medical staff, mostly people that are unfamiliar with body-worn inertial sensors, tend to interpret basic time-series visualizations from an activity recognition system for monitoring patients. To make our study as realistic as possible, we gathered the inertial data and according visualizations from a set of 33 *patient* participants that were unknown to the 15 *medical staff* participants that were shown the visualizations. The predictions of activities from the *patient* set data were obtained from the explained detection system. Figure 5 illustrates the main steps of the evaluation. The following describes the basic steps taken for the experiment in more detail:

- We recorded inertial data from 33 participants, our *patient* data set, that each wore a wrist-worn data logger for 5

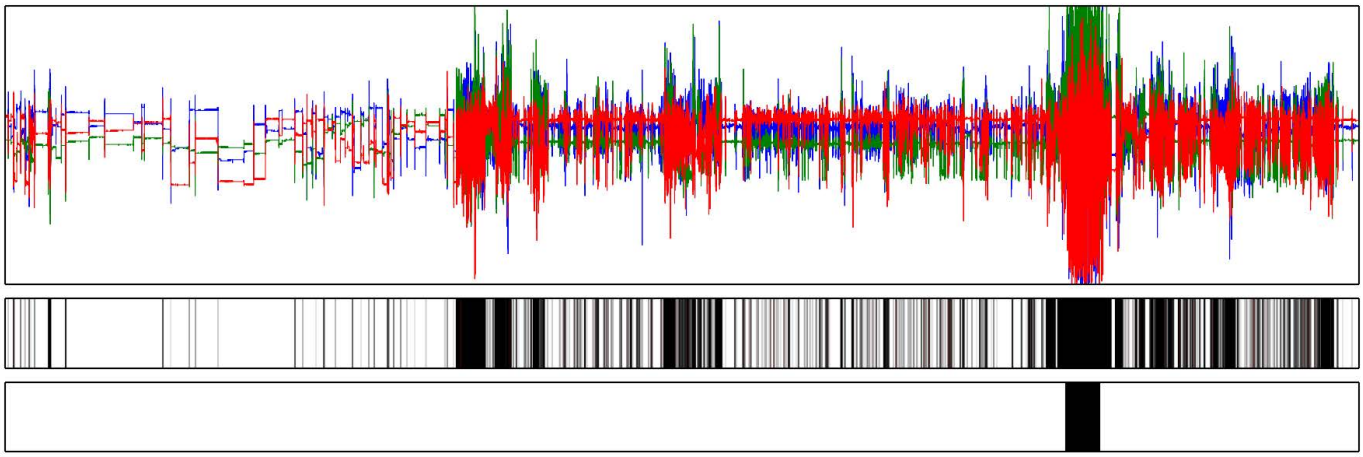

Figure 4. The three basic visualization types we have investigated as feedback to the medical staff examining a patient's activity data: (top, A) raw inertial data from a 3D accelerometer, (middle, B) motif occurrences for an activity, and (bottom, C) classification of an activity. In this figure we used the example of a person performing the activity badminton.

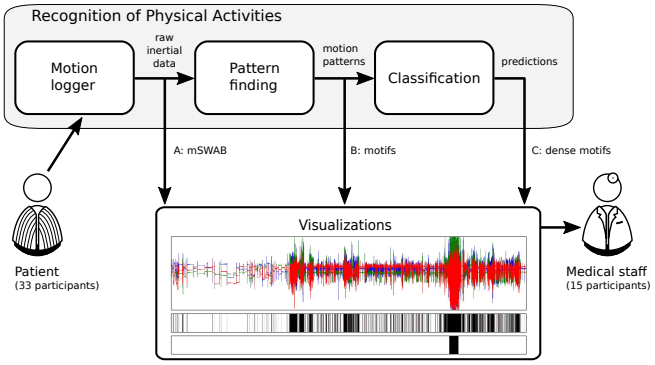

Figure 5. This paper's evaluation uses the data from 33 participants that were monitored for a week each, 24 hours a day, to represent the patients' data (left), which were visualized at three levels of abstraction (bottom plots) to 15 further participants without prior knowledge of inertial data or activity recognition to represent the medical staff (right).

days continuously and performed a regular (usually leisure) activity for about an hour every day.

- This data set was analyzed by our dense motif discovery system and visualizations were made at three different levels of abstraction within the system: (1) raw 3D inertial data, (2) motifs discovered in the data, and (3) the system's predictions for an activity (as per Figure 5).
- 15 participants, generating our *medical staff* data set, with no connection to the 33 *patient* participants, were presented with these visualizations, and were asked to mark where in these a given activity occurred. They were also asked to fill in a questionnaire, e.g., asking about their confidence in each of their estimates.

Figure 4 shows an example of the visualization of these three levels over 24 hours, which in this case includes the activity badminton. The three axes of the raw data are shown over time in the top plot, the middle plot contains occurrences in time of the motifs (as markers), and the predictions of the system are shown in the bottom plot (black boxes). In the study, one of these visualizations is either shown individually or in combination with one or two others, leading to 7 possible combinations of visualizations.

For obtaining realistic performance figures for the activity recognition, the dataset from each participant was split into separate blocks of about a full day (24 hours $\pm$50 minutes) each to facilitate 5-fold cross validation. Each activity instance, depending on the actual activity, generally lasted between 30 and 90 minutes, except for the *fishing* activity, where the activity instances lasted approximately 4 hours each. The target activities that were to be recognized by the system within the data on average consist of $\pm$5% of the entire day, with the rest being other daily activities (details in Table 1).

The *Patient* Data Set

The long-term *patient* data used in the following visualization experiment comes from 33 volunteers or whom a regular physical leisure activity was known before the recording phase, which they would do once each day, over the course of a whole working week. We argue that this closely resembles the type of data that would be gathered by a psychiatric patient over the course of several weeks in-between interviews, although in our data set all participants were volunteers with no known psychiatric disorders, and with no known connections to any participants in the *medical staff* set. For most, this turned out to be a leisure activity or sports, for some a household related activity that was part of their daily schedule. The data was logged by an open-source wrist-worn sensor as introduced before, recording 3D acceleration at 100 Hz on a local microSD card and worn continuously for a working week.

Table 1 gives an overview of all participants who wore our sensor day and night for about a week, specifying their gender, age and the chosen target activity which will be used for testing the detection accuracy of the chosen approaches. The performance of the system, represented in the common performance measurement methods lays at an average of 76% (using the F1 detection rate measure).

Table 1. For every participant in the experiment's *patient* dataset, one physical target activity was chosen prior to the study, to be performed once every day of the week. The activity recognition performance of the dense motif discovery approach is given by precision, recall, and F1-score (in %).

subj.	gender	age	target activity	precision	recall	F1
1	male	30	badminton	95.2	90.0	92.5
2	male	32	badminton	94.4	89.6	91.9
3	male	31	basketball	96.2	92.4	94.3
4	female	26	canoeing	82.5	76.5	79.4
5	male	32	cooking	38.9	42.0	40.4
6	male	35	cycling	87.8	89.0	88.4
7	male	30	dancing	92.2	92.0	92.1
8	female	14	dancing	84.6	86.4	85.5
9	female	16	dancing	44.9	63.5	52.6
10	male	20	drums	91.5	97.6	94.4
11	male	31	fishing	62.6	77.0	69.1
12	male	53	fishing	47.0	81.1	59.5
13	female	26	flamenco	62.7	61.1	61.9
14	male	27	guitar	83.7	79.8	81.7
15	female	27	guitar	94.2	91.0	92.6
16	male	23	guitar	77.6	73.3	75.3
17	male	28	gym	61.6	67.6	64.5
18	male	32	gym	86.1	77.5	81.6
19	male	30	gym	59.6	62.7	61.1
20	female	28	gym	76.4	52.2	62.0
21	male	31	ironing	93.4	84.5	88.7
22	female	27	keyboard	91.7	85.6	88.6
23	female	28	knitting	58.8	73.7	65.4
24	male	30	lunch	26.7	30.1	28.3
25	male	25	soccer	98.1	93.2	95.6
26	female	25	squash	92.0	74.1	82.1
27	male	27	squash	90.4	77.4	83.4
28	male	29	streetdance	66.4	65.8	66.1
29	female	30	streetdance	58.4	69.1	63.3
30	male	32	washing car	81.3	79.9	80.6
31	female	28	xbox	95.5	96.2	95.9
32	female	28	yoga	66.6	43.2	52.4
33	female	30	zumba	96.8	97.6	97.2
			average	76.8	76.1	76.0

Study Methodology

In total, 15 participants (5 male, 10 female) took part in the study, evaluating the visualizations based on the results of the dataset. The participants were chosen to fit as closely to the medical staff scenario as possible, with a non-technical background and no experience in interpreting activity data. Most of the participants were recruited at the university, at non-engineering faculties, with their age varying between 23 and 57 (mean 32) years old. During the experiment, each participant was shown one of the visualizations for an activity which was chosen randomly out of the 33 leisure activities from the *patient* dataset. The order of showing the types of visualization was divided into three phases: (1) the first includes only single visualizations, (2) the second all combinations of two, and (3) the third all visualizations combined. Inside of the phases, the order was also randomized, whereby overall every type was shown twice during each participant's experiment (resulting in 14 visualization iterations per participant). We chose the two runs to increase the amount of data without overloading the concentration of the participants.

At the start of the study, each participant was briefly introduced to the types of visualizations and was given basic information on the underlying abstractions, including the fact that the system's detection rate is on average 76% for any activity in the data set. The information about the detection rate should give the participants a potential hint how to evaluate the system's performance.

The main task for each participant was to estimate where in a given visualization a certain activity had taken place. The name of the activity and the average duration were provided; The participants had to draw in the interval on the printed out visualizations. The participants' estimates were evaluated by comparing the given interval times to the ground truth provided in the data set: if both had an overlap of at least 50%, the estimate was accepted as right, if not, it was rejected as wrong. We use an overlap threshold of 50% as it is commonly used in computer vision object detection (e.g., as used in [11]). Additionally, the participants were asked how certain they were with their estimation and about how intense they thought this activity was performed, using a 7-point Likert-scale.

After the study participants had interpreted all 14 visualizations in this manner, they were asked to answer additional questions on how helpful they found the different combinations, and how much they trusted their judgments based on the different types of visualization. For both, they were asked to rank the three best-performing (combinations of) visualizations. As a final question, they were asked to characterize the activities and their ability to successfully estimate the interval where the activity had taken place.

Results

We start with the quantitative results, more specifically the evaluation of how well the study participants were able to estimate where the activity happened in each visualization and how confident they were in doing so. In addition we also show some qualitative results, where the focus was put on investigating how participants interpreted the different types of visualizations, both based on the study questionnaires, well as on the experimenters' observations made during the study and noted down in a study protocol.

The following are the overall quantitative results for the estimation of activities and participants' confidence therein:

Visualizing the activity recognition system's prediction alone leads to the best overall accuracy: The average performance of the participants' estimation of where the activity took place for the different types of visualization is depicted in Figure 6. The barplot shows that the estimations made were most accurate when only the system prediction (C) is presented to the participant. This was followed by the combination of all three visualizations (A+B+C) and motifs in combination with the system predictions (B+C). The estimation performance on the raw data visualization (A) is particularly poor.

Confidence levels similar to accuracy but with slight differences: When asked about how confident they were in their estimate, the participants showed a similar trend: Figure 7 shows the average participant confidence in their estimate, per combination of visualizations (from low, 1, to high, 7). The self-evaluation of the participants was good, because the predictions alone and the motifs plus predictions were here the top three favorites, though it is interesting to note that the

combination of all visualizations made the participants the most confident overall (even when the difference is too little to be seen as significant in relation to the amount of examined data).

Participants appreciate raw data and motifs: After obtaining the estimates for the 14 chosen activities, the participants were asked to provide a ranking of the three best visualizations. Figures 8 and 9 show the preferred visualizations and their ranking for all participants as a stacked bar visualization. The outcome is as follows: The raw data plus predictions (A+C) and the combination of all three (A+B+C) are the most helpful visualizations, followed by the raw data plus motifs (A+B). It is interesting to note here that what the participants marked as helpful is often contrary to the results of the estimate performance and confidence. Most noticeable are the predictions seen as least helpful, although these had the best performance in using this visualization. The types of visualization which include raw data (A, A+B, A+C, A+B+C) have in both questions reached the highest ratings. Many participants also mentioned in the comments section of the questionnaire that they preferred the inclusion of the raw data visualization.

On a qualitative level, some interesting aspects could be observed as well, regarding how the participants worked during the experiment. We were specifically interested in how the study data was evaluated and how participants were able to reflect on the different levels of the automated recognition system. The last question of the questionnaire was asking how the activities could be characterized which had to be answered in free-form by the participants. The responses give a good insight in the participants' understanding of the data and the activity recognition system.

Intensive physical activities are easiest to grasp: The most common answer was that they can detect sport activities (6) best. The answers mentioning the intensity of activities (4) or fast movements (2) also point in the same direction. Only two study participants turned out to be very different by saying that they can detect best activities with no continuous movements (1) or saying that they can detect movements with very different orientations of the movements (1). One participant pointed out that she has not at all considered activities or their characteristics, but was just orienting herself on the patterns in the visualizations shown to her.

Time of day would help: Equally interesting was the fact that many study participants were thinking about the time of the day a lot while forming their estimations. The time axis on every graph was slightly randomized and had no timescale, so the only possibility to use knowledge about the time of the day is to recognize the sleeping activities in the visualization, which can be done implicitly from the raw data visualization. Two participants mentioned this explicitly and interpreted, based on approximate times, that they were especially confident in detecting activities which are normally done in the evening.

Discussion of the Results

The most significant and interesting result of our study is that there is a big difference between the estimation results and the perception of what the participants think is helpful for them.

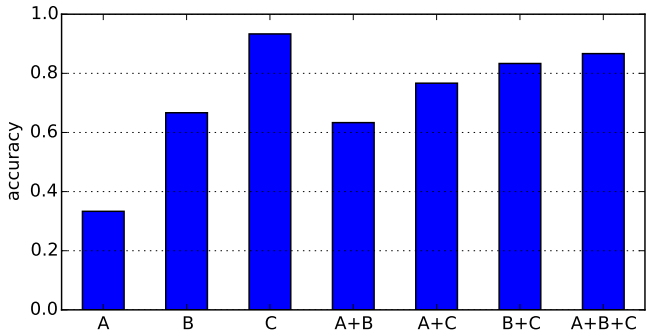

Figure 6. The participants' overall accuracies in estimating when an activity occurred, given (a combination of) visualizations as per Fig. 3.

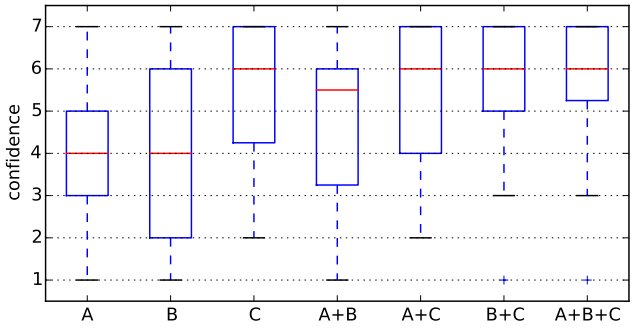

Figure 7. The participants' confidence levels (7-point Likert scale) assigned by themselves on each of their estimates of when an activity occurred.

We can also find a similar difference from the estimation results to the trust, participants have in the visualizations as well (less clear) to the participant's confidence on their estimations. There is a correlation between confidence, trust and helpfulness from which we do not know how the participants have specified it in their evaluation.

To be more precise, adding raw data or motifs to the prediction visualization for the user provides no benefit in the quality of the users' estimations. The results of the study show that independently from that the participants feel more comfortable in their estimates, as well as are thinking of the additional information as more helpful and trust the visualizations even more.

That the quality of the estimation is decreasing seems odd at first because there is extra information added. Though this can be explained by the fact that some participants, when at some point distrusting the recognition system, might get stronger hints at estimating something else when motifs or raw data are provided. This means, even when users do not trust the results of the recognition system, but are confronted with only the systems estimations (C), they have only the change to belief the recognition system is right this time or guess something fully random. When, instead, the system additionally provides insight in raw data or motifs, the users start thinking about it and may guess more often something different from the system's prediction what was in average more often wrong than just following the prediction of the system. This could be an argument for not providing additional visual information,

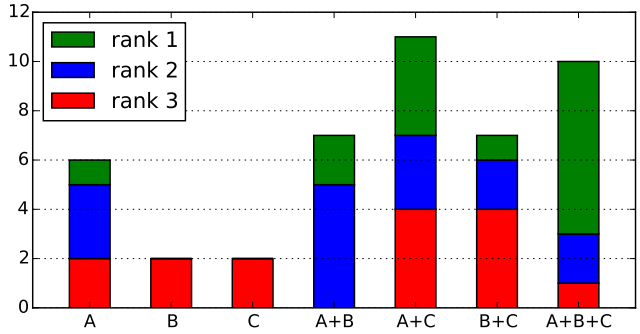

Figure 8. The top three visualizations mentioned by the study participants as being the most *helpful* in pin-pointing the physical activity.

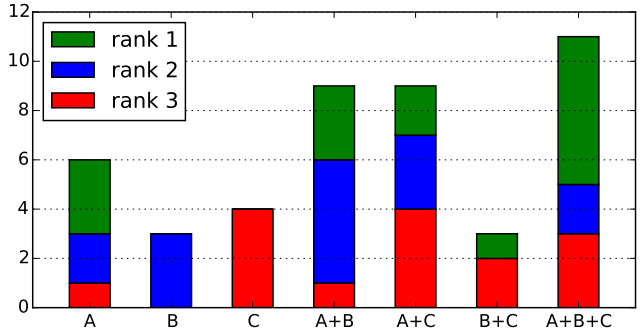

Figure 9. The top three visualizations mentioned by the study participants as the most *trustful* to rely upon when estimating when the activity happened.

but we think that the fact of users feeling more confident and having more trust, which would lead to a better user acceptance of such a system, should be considered as important.

The trustfulness increases especially when the users have access to the original acceleration data (raw data visualization). Having a visualization of motifs as the intermediating stage can help to get an indicator of how the system works and reaches it results. Reflecting on this was not as important than having the raw data though.

What we consider to be very interesting to investigate in future work is the influence of training effects. In this study we have tried to avoid training effects to keep the set of influencing factors small (which was done by randomizing the order and the data presented, although this could have still happened over several visualization runs). With respect to the described scenario, we would assume that with training of doctors the additional information (like raw data) can help raising trust and convenience without the quality downside. It may help detecting misclassification of the system or assessing the conditions of a patient more accurately when, for example, having an indication on the the intensity of activities based on the raw acceleration data.

CONCLUSIONS AND FUTURE WORK

What happens when engineers devise an activity recognition system that can track what activities psychiatric patients are performing over the course of several weeks? This paper has investigated how laypeople would be able to use such a system, with a particular focus on what information to visualize: (A) the raw inertial data, (B) the detected motifs, or (C) the classification estimated from the system.

We have performed a study with a state-of-the-art activity recognition system and evaluated it on a large dataset of 33 participants that recorded their inertial wrist data over a week. Under constraints similar to that of psychiatry monitoring, we asked 15 additional participants without ties to the aforementioned 33, to estimate when an activity was performed, using combinations of the three (A, B, and C) visualizations.

Instead of following a visualization strategy which is independent from the recognition system design, the chosen approach follows a design strategy which reflects the detection system. The visualization includes three levels based on the detection process that includes, besides the results of the used activity classifier, an approximation of the original data on which the detection is based (raw data) and a glance at the "inner" functions of the detection system (motifs, or characteristic motion patterns). The detection accuracy is state of art relative to the high amount of data that is gathered. As any such activity recognition system, it is not perfect, however, and there remains thus a responsibility on the interpreter's side to trust or not to trust the results.

Our user study has shown several interesting results. Having additional visualizations along the classification, such as also presenting the raw data or data from detection steps, is not always a guarantee that people will be able to read the activity data better. In fact, participants performed slightly better when they were presented with just the results from the classifier.

On the other hand, most people especially liked the presence of the raw and intermediate data visualizations and the results showed that these additions helped them trust the data more. Having a means to look at the activity detections at a lower level, therefore might be especially helpful when confidence in the system's prediction is low (e.g., when the system does not detect a particular activity, but the patient insists she performed it that day). The possibility for the medical staff to reflect on activity data on different levels of the detection system is therefore promising and has clear advantages for a trustful use and interpretation of the system's results.

REFERENCES

1. O. Amft, H. Junker, and G. Troster. 2005. Detection of eating and drinking arm gestures using inertial body-worn sensors. In *Proceedings of the 2005 International Symposium on Wearable Computers (ISWC '05)*. 160–163.

2. P. Bech, T. G. Bolwig, P. Kramp, and O. J. Rafaelsen. 1979. The Bech-Rafaelsen Mania Scale and the Hamilton Depression Scale. *Acta psychiatrica Scandinavica* 59, 4 (Apr 1979), 420–430.

3. Eugen Berlin and Kristof Van Laerhoven. 2012. Detecting Leisure Activities with Dense Motif Discovery. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12)*. ACM, 250–259.

4. CamNtech. 2015. MotionWatch 8. (2015). http://www.camntech.com/products/motionwatch/motionwatch-8-overview Accessed: August 2015.
5. K.-H. Chang, M. Y. Chen, and J. Canny. 2007. Tracking free-weight exercises. In *Proceedings of the 9th International Conference on Ubiquitous Computing (UbiComp '07)*. Springer-Verlag, 19–37.
6. T. Choudhury, S. Consolvo, B. Harrison, and others. 2008. The Mobile Sensing Platform: An Embedded Activity Recognition System. *IEEE Pervasive Computing* 7, 2 (2008), 32–41.
7. Sunny Consolvo, David W. McDonald, Tammy Toscos, Mike Y. Chen, Jon Froehlich, and others. 2008. Activity Sensing in the Wild: A Field Trial of Ubifit Garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, 1797–1806.
8. P. Corkum, R. Tannock, H. Moldofsky, S. Hogg-Johnson, and T. Humphries. 2001. Actigraphy and parental ratings of sleep in children with attention-deficit/hyperactivity disorder (ADHD). *Sleep* 24, 3 (May 2001), 303–312.
9. Andrea Cuttone, MichaelKai Petersen, and JakobEg Larsen. 2014. Four Data Visualization Heuristics to Facilitate Reflection in Personal Informatics. In *Universal Access in Human-Computer Interaction. Design for All and Accessibility Practice*, Vol. 8516. Springer Intl' Publishing, 541–552.
10. Daniel Epstein, Felicia Cordeiro, Elizabeth Bales, James Fogarty, and Sean Munson. 2014. Taming Data Complexity in Lifelogs: Exploring Visual Cuts of Personal Informatics Data. In *Proc. of the 2014 Conference on Designing Interactive Systems (DIS '14)*. ACM, 667–676. DOI: http://dx.doi.org/10.1145/2598510.2598558
11. Mark Everingham, Luc Van Gool, ChristopherK.I. Williams, John Winn, and Andrew Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* 88, 2 (2010), 303–338. DOI: http://dx.doi.org/10.1007/s11263-009-0275-4
12. Raffay Hamid, S. Maddi, A. Bobick, and I. Essa. 2007. Structure from Statistics - Unsupervised Activity Analysis using Suffix Trees. In *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*. 1–8.
13. T. Holleczek, J. Schoch, B. Arnrich, and G. Troster. 2010. Recognizing turns and other snowboarding activities with a gyroscope. In *Proceedings of the 2010 International Symposium on Wearable Computers (ISWC '10)*. 1–8.
14. Kevin Huang, Patrick J. Sparto, Sara Kiesler, Asim Smailagic, Jennifer Mankoff, and Dan Siewiorek. 2014. A Technology Probe of Wearable In-home Computer-assisted Physical Therapy. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, 2541–2550. DOI: http://dx.doi.org/10.1145/2556288.2557416
15. T. Huynh and B. Schiele. 2005. Analyzing features for activity recognition. In *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies (sOc-EUSAI '05)*. 159–163.
16. S. Intille, K. Larson, E. Tapia, and others. 2006. Using a live-in laboratory for ubiquitous computing research. In *Proceedings of the 4th International Conference on Pervasive Computing (PERVASIVE' 06)*. Springer-Verlag, 349–365.
17. Rohit Ashok Khot, Jeewon Lee, Deepti Aggarwal, Larissa Hjorth, and Florian 'Floyd' Mueller. 2015. TastyBeats: Designing Palatable Representations of Physical Activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, 2933–2942. DOI: http://dx.doi.org/10.1145/2702123.2702197
18. Jakob Eg Larsen, Andrea Cuttone, and Sune Lehmann Jørgensen. 2013. QS Spiral: Visualizing Periodic Quantified Self Data. In *Proceedings of CHI 2013 Workshop on Personal Informatics in the Wild: Hacking Habits for Health and Happiness*.
19. J. Lester, T. Choudhury, N. Kern, G. Borriello, and B. Hannaford. 2005. A hybrid discriminative/generative approach for modeling human activities. In *Proceedings of the 19th international joint conference on Artificial intelligence (IJCAI '05)*. 766–772.
20. James J. Lin, Lena Mamykina, Silvia Lindtner, Gregory Delajoux, and Henry B. Strub. Fish'N'Steps: Encouraging Physical Activity with an Interactive Computer Game. In *Proceedings of the 8th International Conference on Ubiquitous Computing (UbiComp '06)*. Springer-Verlag, 261–278.
21. M. Mahdaviani and T. Choudhury. 2008. Fast and Scalable Training of Semi-Supervised CRFs with Application to Activity Recognition. In *NIPS '07*. MIT Press, 977–984.
22. D. Minnen, T. Starner, I. Essa, and C. Isbell. 2006. Discovering Characteristic Actions from On-Body Sensor Data. In *Proceedings of the 2006 International Symposium on Wearable Computers (ISWC '06)*. 11–18.
23. Andreas Möller, Matthias Kranz, Barbara Schmid, Luis Roalter, and Stefan Diewald. 2013. Investigating Self-reporting Behavior in Long-term Studies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, 2931–2940.
24. N. Oliver and E. Horvitz. 2005. A comparison of HMMs and dynamic bayesian networks for recognizing office activities. In *Proceedings of the 10th international conference on User Modeling (UM '05)*. Springer-Verlag, 199–209.

25. J. Partridge, K.; Begole. 2009. Activity-based Advertising: Techniques and Challenges. In *Workshop on Pervasive Advertising*.

26. D. J. Patterson, D. Fox, H. Kautz, and M. Philipose. 2005. Fine-grained activity recognition by aggregating abstract object usage. In *Proceedings of the 2005 International Symposium on Wearable Computers (ISWC '05)*. 44–51.

27. M. Philipose, K. Fishkin, M. Perkowitz, D. Patterson, D. Fox, H. Kautz, and D. Hahnel. 2004. Inferring activities from interactions with objects. *IEEE Pervasive Computing* 3, 4 (2004), 50–57.

28. T. Stiefmeier, D. Roggen, G. Ogris, and P. Lukowicz. 2008. Wearable Activity Tracking in Car Manufacturing. *IEEE Pervasive Computing* 7, 2 (2008), 42–50.

29. D. Tacconi, O. Mayora, P. Lukowicz, B. Arnrich, G. Tröster, and C. Haring. 2007. On the feasibility of using activity recognition and context aware interaction to support early diagnosis of bipolar disorder. In *Ubiwell Workshop '07*. 206–209.

30. M. H. Teicher. 1995. Actigraphy and motion analysis: new tools for psychiatry. *Harvard review of psychiatry* 3, 1 (Jun 1995), 18–35.

31. A. Vahdatpour, N. Amini, and M. Sarrafzadeh. 2009. Toward unsupervised activity discovery using multi-dimensional motif detection in time series. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence (IJCAI'09)*. 1261–1266.

32. K. Van Laerhoven, E. Berlin, and B. Schiele. 2009. Enabling Efficient Time Series Analysis for Wearable Activity Data. In *Proceedings of the 8th International Conference on Machine Learning and Applications (ICMLA '09)*. IEEE, 392–397.

33. K. Van Laerhoven, H.-W. Gellersen, and Y. G. Malliaris. 2006. Long term activity monitoring with a wearable sensor node. In *Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks (BSN '06)*. 174–177.

34. F. Wilhelm, M. Pfaltz, and P. Grossman. 2006. Continuous electronic data capture of physiology, behavior and experience in real life: towards ecological momentary assessment of emotion. *Interacting with Computers* 18, 2 (2006), 171–186.