Smartphone-Based Recognition of States and State Changes in Bipolar Disorder Patients

Agnes Grünerbl, Amir Muaremi, Venet Osmani, Gernot Bahle, Stefan Öhler, Gerhard Tröster, Oscar Mayora, Christian Haring, and Paul Lukowicz

***Abstract*—Today's health care is difficult to imagine without the possibility to objectively measure various physiological parameters related to patients' symptoms (from temperature through blood pressure to complex tomographic procedures). Psychiatric care remains a notable exception that heavily relies on patient interviews and self-assessment. This is due to the fact that mental illnesses manifest themselves mainly in the way patients behave throughout their daily life and, until recently there were no "behavior measurement devices." This is now changing with the progress in wearable activity recognition and sensor enabled smartphones. In this paper, we introduce a system, which, based on smartphone-sensing is able to recognize depressive and manic states and detect state changes of patients suffering from bipolar disorder. Drawing upon a real-life dataset of ten patients, recorded over a time period of 12 weeks (in total over 800 days of data tracing 17 state changes) by four different sensing modalities, we could extract features corresponding to all disease-relevant aspects in behavior. Using these features, we gain recognition accuracies of 76% by fusing all sensor modalities and state change detection precision and recall of over 97%. This paper furthermore outlines the applicability of this system in the physician–patient relations in order to facilitate the life and treatment of bipolar patients.**

***Index Terms*—Activity recognition, bipolar disorder, depression recognition, mental disease monitoring, mood recognition, smartphones, wearable computing.**

I. Introduction

A. Activity Recognition and Mental Disorders Treatment

BIPOLAR disorder [1] is a common and severe form of mental illness. People suffering from this disorder experience more or less frequent successions of periods of manic, normal, and depressive state. The current standard for determining the severity of an episode uses subjective clinical rating-scales based on self-reporting that were developed in the early 1960's (e.g., HAMD, BRAMS scales) or more recent variations of them (e.g., BSDS). While the efficacy of these scales has been proven, they still are a potential source of subjectivity and additionally require the attendance of a trained professional. The main treatment currently offered is a life of pharmacotherapy, which has to be modified according to a patient's state. Additional substances may have to be prescribed to increase the prophylactic effect of the therapy. Even so, the effectiveness of treatment strongly depends on the timing. Thus, therapeutic measures can be very effective if administered at the beginning of a patient's transition into a different state (e.g., from normal to depressive). They may be a lot less effective if severe symptoms have persisted for a significant time. As a consequence, a promising form of intervention is teaching patients to recognize and manage early warning signs (EWS). A systematic review of this approach found that 11 randomized controlled trials (RCTs) involving 1324 patients show the efficacy of interventions that include EWS self-recognition [2]. However, this involves a very significant training effort (which is difficult to finance) and strongly depends on the patients' compliance and discipline. Thus, in some cases, it can be impractical or even impossible and therefore it's usage has limitations.

Cognitive, mental, and emotional disorders are an obvious application field for activity recognition. As the symptoms of such diseases manifest themselves in changes of behavior [3], activity aware systems could be used as core instruments for assisting diagnosis and treatment. Even more, the fact that psychiatrists currently have few objective and reliable alternatives would amplify the value of such a system. Ever since X-rays have become available, it is much easier to see exactly how extensive a fracture of a broken limb is and how best to attend it. On the contrary, most of the time psychiatrists have to rely on a patient's subjective recollection of their behavior. The closest thing to a "measurement" are self-assessment questionnaires that can be time consuming and rely on subjective recollections and the patients' self-perception only. As a consequence patients often end up visiting the doctor very late, which makes treatment more difficult and often leads to the necessity of severe measures and prolonged hospitalization. On one hand, this can have a dramatic impact on the patient's life (long sick leaves) and on the other hand is of costly relevance to the health system.

While the benefit of a more "objective" measurement based on activity recognition is clear, developing and implementing such a system is difficult for many reasons. First, having people suffering from a mental disorder wear complex sensors on a daily basis is often not practicable. Second, since there are no reliable automatic diagnostic instruments, getting enough ground truth for training and testing involves a huge effort in terms of long

Manuscript received March 31, 2014; revised July 8, 2014; accepted July 17, 2014. Date of publication July 25, 2014; date of current version December 30, 2014. This work was supported by the MONARCA Project (www.monarca-project.eu) from the EU FP7.

A. Grünerbl, G. Bahle, and P. Lukowicz are with the Department of Embedded Intelligence, German Research Center for Artificial Intelligence, 67663 Kaiaserslautern, Germany (e-mail: agnes.gruenerbl@dfki.de; gernot.bahle@dfki.de; paul.lukowicz@dfki.de).

A. Muaremi and G. Tröster are with the Wearable Computing Lab, ETH Zurich, Zurich CH-8092, Switzerland (e-mail: muaremi@ife.ee.ethz.ch; troester@ife.ee.ethz.ch).

V. Osmani and O. Mayora are with the Mobile and Ubiquitous Technology Group, CREATE-NET, 38123 Trento, Italy (e-mail: venet.osmani@create-net.org; oscar.mayora@create-net.org).

S. Öhler and C. Haring are with the Department for Psychiatry and Psychotherapy B, State Hospital, 6060 Hall in Tyrol, Austria (e-mail: stefan.oehler@tilak.at; chrisitan.haring@tilak.at).

Digital Object Identifier 10.1109/JBHI.2014.2343154

running trials involving repeated appointments with professionals. Finally, the fact that behavior can vary strongly on a daily basis, independently of illness-based effects, makes recognition difficult. As a consequence, very little work exists on diagnostic work using pervasive sensors in real-world environments.

By overcoming such difficulties, we demonstrate in this paper how smartphone usage patterns and sensor data can be used as an objective "measurement device" for aiding psychiatric care. By drawing upon real-life data of ten bipolar patients (more than 800 sensor traces), we show how the recognition of the mental state of the patients and the detection of changes therein can be accomplished to a high degree of accuracy. We build upon previous work, significantly improving the entire system by combining more, different sensor modalities and therefore including different disease-relevant aspects of human behavior. We achieve a state recognition accuracy of 76% (with low variance, ranging from 68% to 81%). For the change of mental state detection, we obtain precision and recall of about 97%.

B. Related Work

1) Self-Assessment of Mental Disorder Symptoms Using Mobile Phones: Smartphones are playing an increasingly important role in gauging mental health. For instance, there are apps designed for self-assessment that can help patients to estimate and monitor symptoms specific to their ailment, which can then be shared with psychiatrists. For example, the eMoods Bipolar Mood Tracker app [4] provides a system that allows users to input subjective mood ratings daily and monitor them via an electronic journal. The app can also keep track of hours of sleep, anxiety levels, and medication use, which are all self-reported, and can be shared with a family member, caregiver, or clinician.

A number of other approaches have looked at incorporating Ecological Momentary Assessments in order to gather patient state at opportune times [5] specifically for anxiety and eating disorders [6] and also provide Ecological Momentary Interventions. In this study, authors stress the use of external context clues, based on sensor data such as location and social interaction, to deliver effective interventions. Another set of studies that relied on self-monitoring of patients with severe mental illness, specifically bipolar disorder and schizophrenia are presented in [7] and [8]. Authors report evidence of short-term adherence to and acceptability of mobile devices, while emphasizing that it is likely impractical for patients to respond to daily surveys ad infinitum, stressing that context-awareness of mobile devices and sensor sampling can provide feedback relevant to detected patterns of behavior. Similarly, an RCT [9] revealed that while self-reporting and self-assessment of patient state has a positive effect in increasing emotional self-awareness in patients suffering from depression, anxiety, and stress, the mental health outcomes did not improve significantly. As such, considering the impracticality of this method for long-term monitoring [7] and patients' reluctance to log information [10], there is a clear need to infer patients' states in an autonomous manner. Recent trends in this area have been pointing toward using sensors on smartphones for passive collection of objective information.

2) Objective Monitoring of Symptoms of Mental Disorders: Objective monitoring consists of smartphone sensors passively collecting data that can be used to infer patient state. Matthews *et al.* [11] outline different aspects in balancing sensing and patient's need and describe MoodRhythm, a system for tracking daily rhythms. There is far less work in automatically inferring patient state in comparison to self-reported information. One possible approach is to develop systems that predict patient state by using predefined algorithms that are initialized based on evidence from scientific or clinical knowledge [12], [13]. This has been the typical approach of systems that recognize patient activities, where algorithms make inferences regarding the patient's status by plugging in sensor data.

3) Wearable Technology in Health Care: The usage of wearable and pervasive technology in health-care has already been explored in numerous publications in previous years. Overviews include [14] and [15]. Specific examples range from early works about assisting the elderly with cognitive impairment [16], to more recent works about monitoring children's developmental progress using augmented toys and activity recognition [17]. In the area of mental health, the majority of systems deployed to date focus on supporting self-monitoring. Systems that provide patient feedback through questionnaires or text messages are analyzed in [18] and [19].

Other systems, like [20]–[22] similarly rely on self-reporting, implemented using smartphone applications. Burns *et al.* [20], for instance, introduce an app for mood prediction of depressive patients. However, it requires constant interaction and feedback of the patient. Furthermore, "TrueColours" [21] and the "Optimism App" [22] were developed to log self-reported mood, activities, and quality of sleep in order to monitor depression and state changes. LiKamWa *et al.* [23] also display an approach, which infers mood through analyzing mobile phone usage.

4) Automatic Recognition of Mental State: In terms of automatic recognition of mental state much less work exists, in particular work involving real-world studies and off-the-shelf devices like smartphones. In [24] and [25], the usage of an indoor location system to assess the state of dementia patients is presented. Massey *et al.* [26] describe an experimental analysis of a mobile health system for mood disorders where they introduce different possible sensors for mood detection, yet focus on technical aspects like line of sight and reception rate, optimal coverage and optimal placement of on-body sensors. Two publications close to the work presented in this paper are the research done by a group from Denmark [27] and the previously mentioned [20] that introduces a mobile phone application which employs machine learning models to try to predict patients' mood (of depressive patients). Here, however, the ground-truth is fully self-rated, no objective psychological or psychiatric assessment is performed.

In [27], Frost *et al.* use a self-developed smartphone application to record subjective and objective data from patients suffering from bipolar disorder. Even though their main focus lies on self-reported information, in passing they also utilize coarse objective sensor data (acceleration fragments and phone call statistics) to try to estimate future shifts of a patient's mental

state. These predictions are then compared to forecasts derived from the self-reporting data. By contrast, our work goes into far more depth in the area of state classification, also uses location sensors and sound in addition to acceleration and instead of social interaction sensing compares the results to an objective, diagnostic ground-truth on a day-to-day basis. In previous works, our group has also discussed the basic concepts of using smartphones for the management of bipolar disorder [28] and used a smaller (six patients) dataset from a preliminary experiment to detect correlation between selected sensor data and self-reported state (see [29] and [30]).

II. Study Overview and Data Acquisition

A. Vision—Activity Recognition Assisting Mental Care

After having numerous discussions with psychiatrists (see also [28]) and other health care providers, we were able to design a practical and utilizable collaboration of activity recognition and mental care. Its aim has been to develop an application based on smartphone behavior and activity monitoring, usable as an "objective" measurement that helps to detect state changes in order to guarantee the availability of in-time treatment. More specifically, this application should rely solely on objective sensor data and should not require any input/feedback from the user/patient. The last requirement is due to the fact that patients' feedback is very often subjective and carries the risk of being biased. Also, the more interaction asked of the patient, the less compliance can be expected in the long run.

Considering the usability of the envisioned system, some important aspects should be highlighted here:

1) The recognition results of a system as outlined here are not meant to automatically trigger medication. There is no danger that a false recognition would trigger potentially dangerous wrong medication.
2) Required reaction times are on a time scale of a few days rather than a single day. In fact, radical change seldom happens from one day to the next.

Overall, the envisioned usage scenario for the recognition system is to provide daily updates to the doctors and possibly the patients who would then look at the trend evolving on the scale of a few days and, if the trend points toward a negative state change, make sure that an examination is scheduled. This means that for our work:

1) change detection is more important than the recognition of a particular state;
2) therefore, recognition does not need to be perfect to be useful;
3) more important than perfect recognition is that the results are achievable in the real world, in a setting not only realistic but actually real.

This entails genuine patients and no constraints on where and how to wear the phone, nontechnology-savvy users, and irregular availability of data from different sensors.

B. Data Collection Study

In order to develop a system as described previously, capable of working during everyday life of a patient, it was clear that real-life traces from actual bipolar patients rather than artificial, laboratory-derived data would be essential.

Therefore, as a first step, we conducted a challenging real-world study described in previous publications [29], [31]. Here, only an overview shall be given: during ten months, data were collected from ten bipolar patients in a rural area psychiatric hospital in Austria (12 weeks of data recording per patient). The number of patients included in the study was limited by different factors like hospital resources and availability of patients fitting to the inclusion criteria (contractually capable, bipolar disorder diagnosed by ICD 10 classification, age between 18 and 65+; precondition: willingness and ability to deal with modern smartphones). Due to privacy reasons, the selection of patients was entirely done by the ward's psychiatrists and their perception of which patient was capable of dealing with the study's requirements. For the authors of this paper, there was no way to influence this selection except for defining the inclusion criteria (for more detail, see [31]).

To provide ground-truth, psychological state examinations (psychological standard scale tests as HAMD or YMRS) were frequently performed every three weeks (note that more frequent examinations would have biased the output of the state examinations). These examinations resulted in state-grades for each measurement between -3 for severe depression and $+3$ for severe mania with intermediate steps of depression, slight depression, normal (0) slight mania, and mania. In order to "cover" the time between the measurements, specifically trained psychologists spoke to the patients over the phone.

Most of study participants started the data collection in a more or less severe depressive state. Three of the patients started in a manic phase. All of them underwent one or more changes (1–3 per patient) in their mental state during the study. These changes were mainly between two "adjacent" states (e.g., depressive-normal or manic-normal). Overall 17 state changes were recorded.

C. Collecting Data

Each patient was given an Android smartphone running a logging application developed by our group [29], which was designed to record all sensor data automatically in the phone's background. Concerning data recording, no interaction with the patient was necessary. At the end of every day, the patients were asked whether they felt comfortable storing the day's data (and thus providing it to the researchers). If the patient did not agree, all data collected during that particular day would be deleted. Otherwise, it would be stored on the SD card. This protocol was a precondition for the approval of the ethics board. However, during the entire trial, there was no case of a patient asking to delete data. The data from the SD card were copied during the periodic examination and anonymized to hide the patients' identity. Clearly, in a productive system, the data would need to be transmitted wirelessly at the end of each day. However, for the purpose of our research, the SD card option was more reliable and allowed us to simplify data security issues.

D. Data Quality and Ground-Truth

Even though described earlier [31], we want to give some insight about the data quality and amount. Once more it has to be mentioned that the data collection was performed during the everyday life of the patients. This means that the patients could use their phones the way they would normally use it. The downside of this is that patients were sometimes switching off sensors in order to save energy or forgot to charge the battery overnight. As a result, the number of available data is (sometimes significantly) smaller than the recording period of 12 weeks (84 daily datasets per patient) would suggest. Furthermore, for supervised machine learning approaches requiring training, the number of available days is reduced even further. As already explained, ground-truth is only available every three weeks with intermediate telephone interviews (total of 5–9 days per patient). Fortunately, we could draw upon knowledge about patient's behavior and parallel self-assessment in order to extend the ground-truth periods. However, even extended, they do not cover the entire recording period. Thus, the actual amount of sensor data available lies between 19 and 71 datasets per patient per sensor modality. For more details about ground-truth extension and data quality, see [31].

III. Recognizing the State

A. Relevance of Sensors and Features

Initial experiments [29] and several discussions with the medical personnel gave us insight into the relative relevance of different behavioral aspects:

1) *Social interaction:* The way people interact with others can vary quite a lot. What people with bipolar disorder have in common, though, is that in a depressive phase, the desire and ability for social interaction is reduced, while during the manic phase it is heightened.
2) *Physical motion:* Patients with depression tend to move less, move less forcefully and overall slower. The opposite is true for manic patients.
3) *Travel patterns:* Most people have their travel routines dominated by a set of places, which they often visit in a certain temporal pattern. These patterns tend to change in both depressive and manic states (become less frequent or more erratic respectively). In addition, depressive people tend to travel less and be outside less.

Of course, this is to be seen as a statistical average and is strongly person dependent. Some people may move more in depressive state than others do when manic.

B. Social Interaction for State Recognition

Aspects of smartphone usage characteristic of social interaction are, e.g., behavior related to phone calls or text messages. Related to this, and also very indicative of mental state, is sound. For privacy reasons, voice recognition is obviously not an option, but analysis of pitch, talking speed, etc., are very interesting.

1) Feature Extraction: The first step of any standard pattern recognition technique is to extract appropriate features.

a) Phone call features: Phone call behavior that has been recorded includes the length of phone calls, whether phone calls were incoming or outgoing, and which caller ID numbers were involved (note that due to privacy reasons, the numbers were anonymized and only the last four digits were stored in order to identify the counterpart). Using this data, the following features were extracted for each day:

1) number of phone calls;
2) total length of calls (sum of call length per day);
3) average length of phone calls;
4) standard deviation of the length of phone calls;
5) number of unique numbers.

b) Sound features: This paragraph is a summary of the feature sets described in our previous work in [32].

We divide the sound features into speech features which describe the phone call interaction and voice features which are usually used to detect the emotions from the voice.

Speech features: The aim is to understand the dyadic communication of the patient with the other person on the line. Starting from the voice activity detection (voiced speech versus unvoiced speech), the speaking segments are created. Using these segments, we are able to differentiate between turns, short turns, and nonspeaking segments. Short turns or utterances are feedback words while someone else is talking, such as "okay," "hm," "right," etc. Nonspeaking segments are either pauses or turns of the counterpart (see [33] for more details). The following speech features were then calculated on a daily basis:

1) average speaking length and speaking turn duration;
2) average number of speaker turns and short turns/utterances;
3) standard deviation of speaking turn duration;
4) speaker turns per length in minutes and short turns/utterances per length in minutes;
5) percentage of speaking from the total conversation.

Voice features: The open-source "openSmile" toolbox [34] is used to extract the acoustic features. For each frame of the speech signal (frame length: 25 ms, step size: 10 ms), different low-level descriptors are calculated: rms frame energy, mel-frequency cepstral coefficients (MFCC) 1–12, pitch frequency F_0, harmonic-to-noise ratio (HNR), and zero-crossing-rate (ZCR). Then, functionals like mean, standard deviation, extreme values, kurtosis, and more were applied on all frames for each descriptor. The resulting feature vector was reduced by using the filter feature selection method. Finally, we end up with the following voice features:

1) kurtosis energy;
2) mean second and mean third MFCC;
3) mean fourth delta MFCC;
4) max ZCR and mean HNR;
5) std and range F_0.

2) State Recognition Technique: With the features described previously in place, we first attempted to apply standard pattern recognition techniques to the data to try to identify which state a patient had been in. As is common with supervised

TABLE I
ACCURACY IN % (# OF TOTAL INSTANCES) AND RECALL PRECISION FOR PHONE, SOUND, AND SENSOR-FUSION (CLASSES/STATES: SEE SECTION II-B)

Patients	PHONE	Rec/Prec	SOUND	Rec/Prec	Fusion	Rec/Prec
p0102	75(46)	64.4/70.1	66(46)	51.2/40.4	73(46)	58.5/68.0
p0201	62(38)	52.5/53.2	68(32)	60.8/62.0	71(38)	58.7/66.4
p0302	71(60)	62.0/63.6	74(60)	64.5/52.0	71(60)	60.7/65.3
p0602	36(35)	33.9/35.0	76(35)	68.5/78.7	65(35)	57.0/48.0
p0902	68(41)	63.7/65.3	71(41)	68.5/68.5	68(41)	62.3/65.5
p1002	65(37)	78.7/69.4	65(37)	54.0/40.8	65(40)	53.3/41.4
Average	**66%**	**61/58**	**70%**	**60/59**	**69%**	**52/55**

TABLE II
COMPARISON OF RECOGNITION OF DIFFERENT MODALITIES AND FUSION OF DIFFERENT MODALITIES AND OF ALL-IN FUSION

(av. # instances)	**Recall** (std)	**Precision** (std)	**total** (std)
PHONE (43)	54.4% (11.4)	37.3% (29)	64.2% (13)
SOUND (43)	61.3% (7.3)	24.2% (31)	69.8% (4.5)
FUSION P + S (43)	**58.4%** (3.2)	**59.0%** (11.3)	**69.3%** (3.5)
ACC (48)	62.9% (7.9)	64.8% (7.8)	71.7% (3.8)
LOC (33)	72.3% (14.5)	76.5% (13.7)	81.7% (6,5)
FUSION A + L (49)	**66.3%** (11.0)	**57.9%** (16.6)	**75.6%** (6.2)
all-in FUSION (49)	**65.8%** (8.8)	**72.0%** (12.2)	**76.4%** (4.1)

learning methods, we performed a percentage split on our dataset, dividing it into 66% training and 33% test samples. The split was performed randomly. The test set was resampled to ensure that classes were equally represented. For the actual classification, features were first transformed using a linear discriminant analysis [35]. The classes for the classification were defined according to the diagnosis provided by psychologists (depressive, normal, and manic with different degrees—up to seven classes possible); also see Section II-B.

Afterward, the Naïve Bayes classifier included in Weka [36] was used to estimate classes for the test set. Other classifiers were tested (k-nearest neighbor, j48 search tree, conjunctive rule learner), but achieved very similar results. The entire process above was repeated 500 times in a cross-validation approach with random test/training splits to eliminate artifacts caused by "lucky" or "unlucky" random selections.

3) Results: Table I displays the results of the recognition with phone call features and voice (sound). Note that four of the patients refused to use the study phone to make phone calls. Therefore, these patients are not included in the table.

The results are not very satisfying. They show that with plain classification only phone call behavior does not provide a very high recognition rate (only 66% for phone call behavior). Classification with voice features works better (70%), nevertheless even a fusion of both modalities is not able to enhance the result (only 69%). Even though, as pointed out before, classification rates do not have to be perfect to be useful in the outlined context, still, it should be possible to reach better accuracy.

C. Using Other Sensors for State Recognition

Using other sensor modalities like GPS data or acceleration has already proven to work sufficiently [31]. Having extracted different travel patterns and movement features, we could achieve an average recognition accuracy of 70% (acceleration)–80% (location). We could show, furthermore, that a fusion of these two sensor-modalities, even though providing a slightly lower accuracy as location data alone, enhances the reliability of the results. This is due to the well-known fact that classification based on a larger set of data is more reliable than when the dataset is relatively small. In Table II (ACC, LOC, FUSION A + L), a summary of the results of classification (70–80%) and sensor fusion (80%) including acceleration data and location data is given.

D. All-In: Optimizing State Recognition

Summing up, classification with GPS and acceleration data provides better accuracy than simple classification with phone call and sound data. Nevertheless, GPS and acceleration only cover travel patterns and movement, but they do not represent the social behavior aspect at all, which as stated before is a crucial aspect in addressing bipolar disorder.

Only a fusion of all sensor modalities covers the three behavioral aspects described previously. The following steps achieve this.

The previous single sensor classification resulted in a list of probabilities for all possible classes, for each day, for each modality (phone, sound, acceleration, and location). Combining them yielded a final classification for each day data was available for. The combination process was performed as follows: for everyday, where there was only one modality available, the most probable class of the associated class probability list was chosen. For everyday, where more than one modality provided class estimates, those estimates were fused using this algorithm:

1) For each class, the ratio of training data available for acceleration and location compared to all training data was calculated.
2) If modality one provided ten samples of training and modality two provided five samples, the ratio would have been 0.66 for one and 0.33 for two.
3) In order to further penalize little available training data, these coefficients where then input into a sigmoid weighting function: $1/(1 + e^{-(\mathrm{coeff}-0.5)*5})$
4) Finally, the product of estimated class probabilities and coefficients was calculated for each modality. These vectors of class estimates were then summed up and the highest rated class was picked as winner.

Due to the fact that some data modalities were scarce (specifically GPS data), the above scheme was chosen in order to reward results that could rely on a larger amount of data, which is more trustworthy.

In Table II, the result of the fusion and a direct comparison of the recognition and fusion of the different sensor modalities reveals that even though the recognition and fusion of acceleration, and location is better than the recognition and

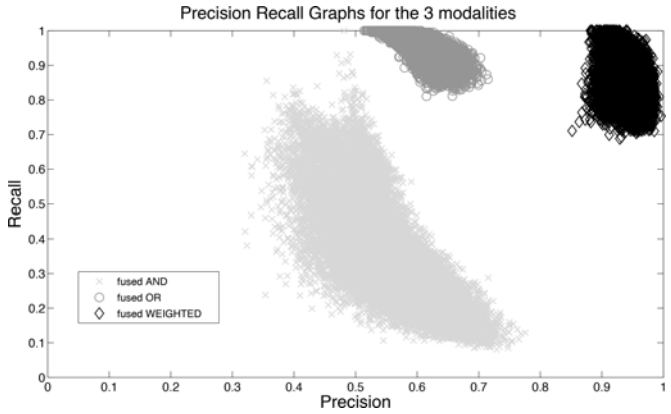

Fig. 1. Precision/recall graph of the three concatenation modalities of the state change detection fusion using different thresholds.

fusion of phone behavior and sound, the all-in fusion including all sensor modalities even increases these results slightly (total of 76% correct classification). This means that, as assumed earlier, the combination of social behavior, travel patterns, and movement, by covering different aspects in the patient's behavior, optimizes the state detection and provides better results than single sensors or fusion of only a few sensor modalities.

IV. Detection of State Changes

Having achieved sufficient recognition results, the next and even more important step is to detect changes in state. In the following section, we will introduce detection of a state change without explicit recognition of the new state. The main idea has already been introduced in earlier work, yet a summary will be given here. For more details, see [31].

A. Method

The difference to the approach in the previous section is that instead of a classifier that has a model for each state relevant to the patient, we only build a model of a single "default state." All points falling outside this model are classified as a "change." In order to determining the border between in the model and outside the model (threshold), a set of values was tested, resulting in the precision/recall graph in Fig. 1. This approach is motivated by the considerations that, from the application point of view, detecting a change of state in order to trigger a visit to the doctor is a key functionality. The exact diagnosis is then done by the doctor anyway. Moreover, the approach of starting with a single default state has the advantage that it is instantly usable as there is no need to wait until data for all relevant states has been collected, and for eight out of ten patients, we are dealing with a two state problem anyway.

B. Optimized State Change Detection

Since in the state change detection case, we are explicitly building our own probability density functions; for the default state, we could use the results of the single sensor state change detection to perform a more controlled fusion strategy where the distances to the mean of the initial state distribution are used as weights to reward/punish results together with the number of available training points.

TABLE III
State Change Det. Results of the Diff. Fusion Modalities

	Recall	Precision
A+L weighted Fusion	**96.40%**	**94.50%**
All-in AND Fusion	42.87%	61.18%
All-in OR Fusion	92.15%	70.28%
All-in WEIGHTED Fusion	**97.36%**	**97.19%**

In order to complete the picture, the fusion of state change detection results for the single sensor modalities was applied in different concatenations of the single modalities (AND, OR, and our own Weighted FUSION). Fig. 1 shows that the recall/precision of the Weighted Fusion concentrates at a corner at the very right top, meaning that the accuracy of this modality provides very high results which are stable for all patients.

The numbers in the result table (see Table III) reveal that an AND concatenation (meaning a state change is detected only when all sensor modalities detect one) is neither very precise nor does it detect many changes (low recall). Obviously, since features come from four distinct areas, fusing by AND would imply that behavior changes significantly in all of them at once, meaning, e.g., a patient has to be outside less, call less, move less, and talk differently. Using an OR concatenation, almost all changes are detected (very high recall) yet there will be a number of falls alarms (lower precision). By applying the self-designed Weighted Fusion concatenation both recall and precision are very high, meaning that almost all changes are detected and almost no false alarms will occur.

Furthermore, in Table III, a direct comparison between the good results of the state change detection with Weighted Fusion concatenation of acceleration and location from our earlier work, and the results of the all-in (all sensor modalities in) Weighted Fusion results is displayed. Once again it is shown by this table that the best accuracy can be achieved in fusing all sensor modalities and therefore all disease-relevant aspects of behavior. Even though the weighted fusion of only location and acceleration data provides very good results, the addition of social behavior, which alone does not provide the best results, clearly enhances the overall accuracy of the state change detection further. This once more underlines, what has been pointed out before, that for a stable recognition and state change detection, it is important to have all behavioral aspects in mind because various aspects together provide the best picture of the actual state of the patient.

V. Conclusion

The work presented in this paper introduces a smartphone sensor-based system dedicated to facilitating the life of bipolar disorder patients and supporting their treatment. It incorporates sensor modalities covering different disease-relevant aspects of the human behavior, but does not rely on self-assessment, thus

providing a less biased and more objective additional information source to health care professionals. The average recognition accuracy of 76% must be seen in light of a noisy ground-truth and the fact that a patient's behavior cannot be expected to be fully consistent on a daily basis, as even a severely depressed person can have a good day. Additionally, our system does not aim to replace professional expertise, but rather supplement it. Considering these aspects, we believe our results to be very promising.

Even more important, at least from a practical perspective, is the early detection of changes in a patient's state. Once again relying on four different sensor modalities, we achieve an almost perfect recall and precision of about 97%. This implies that it is possible to reliably provide early warnings, in turn facilitating less onerous treatments. Moreover, even if a false alarm occurred, it would at most result in an unneeded appointment with a psychiatrist. We explicitly are not aiming for automatic medication or similar notions.

However, the most valuable achievement of this paper is that the introduced system has been derived from and validated by a large, real-world dataset (more than 800 days of sensor recordings), recorded during the everyday lives of real patients. Additionally, we could show that the system can handle irregular availability of data while still providing sufficient results. In our opinion, this is a significant improvement over artificial lab settings and qualitative studies.

Furthermore, a real-world application of the state change detection algorithm is almost plug-and-play as it only requires data of the current state, not all possible states. It can, therefore, be used without excessive training and labeling phases. Basically, the system can aid psychologists from a patient's first visit onward.

To the best of our knowledge, a system like the one introduced here, which is able to detect early changes in the state of a bipolar disorder patient and moreover, works under the constraints of everyday life and does not require long periods of training and calibration is not yet available. For this reason, we believe that the work presented could become a potent tool in supporting the treatment of bipolar disorder.

ACKNOWLEDGMENT

The authors specifically would like to thank the professionals of the psychiatric hospital (psychiatrists and nurses) for their support and patience while the study was conducted.

REFERENCES

[1] National Institute of Mental Health. (2014, Jul.). Bipolar disorder in adults. [Online]. Available: www.nimh.nih.gov/health/publications/bipolar-disorder-in-adults/bipolar_disorder_adults_cl508.pdf.

[2] F. E. A. Lobban, "Enhanced relaps prevention for bipolar disorder: A cluster randomised controlled trial to assess the feasibility of training care coordinators to offer enhanced relapse prevention for bipolar disorder," *BMC Psychiatry*, vol. 7, no. 1, p. 6, 2007.

[3] American Psychiatric Association. (2013, Jun.). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.) [Online]. Available: dsm.psychiatryonline.org

[4] G. Technologies. (2013). Emoods bipolar mood tracker (version 1.0) [mobile software application]. [Online]. Available: https://play.google.com/store/apps

[5] K. Heron and J. Smyth, "Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behaviour treatments," *Brit. J. Health Psychol.*, vol. 15, pp. 1–39, 2010.

[6] J. Treasure, C. Macare, I. Mentxaka, and A. Harrison, "The use of a vodcast to support eating and reduce anxiety in people with eating disorder: A case series." *Eur. Eating Disorder Rev.*, vol. 18, pp. 515–521, 2010.

[7] C. Depp, B. Mausbach, E. Granholm, V. Cardenas, D. Ben-Zeev, T. Patterson, B. Lebowitz, and D. Jeste, "Mobile interventions for severe mental illness: Design and preliminary data from three approaches." *J. Nervous Mental Dis.*, vol. 198, no. 10, pp. 712–721, 2010.

[8] E. Granholm, D. Ben-Zeev, P. Link, K. Bradshaw, and J. L. Holden, "Mobile assessment and treatment for schizophrenia (MATS): A pilot trial of an interactive text-messaging intervention for medication adherence, socialization, and auditory hallucinations." *Schizophrenia Bull.*, vol. 38, no. 3, pp. 414–425, 2012.

[9] S. Reid, S. Kauer, S. Hearps, A. Crooke, A. Khor, L. Sanci, and G. Patton, "A mobile phone application for the assessment and management of youth mental health problems in primary care: A randomised controlled trial." *IEEE Commun. Mag.*, vol. 12, no. 1, p. 131, Nov. 2011.

[10] N. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. Campbell, "A survey of mobile phone sensing." *IEEE Commun. Mag.*, vol. 48, no. 9, pp. 140–150, Sep. 2010.

[11] M. Matthews, S. Abdullah, G. Gay, and T. Choudhury, "Tracking mental well-being: Balancing rich sensing and patient needs," *Comput.*, vol. 47, no. 4, pp. 36–43, 2014.

[12] A. Honka, K. Kaipainen, H. Hietala, and N. Saranummi, "Rethinking health: ICT enabled services to empower people to manage their health," *IEEE Rev. Biomed. Eng.*, vol. 4, pp. 119–139, Nov. 2011.

[13] M. Morris and F. Guilak, "Mobile heart health: Project highlight." *IEEE Pervasive Comput.*, vol. 8, no. 2, pp. 57–61, Apr.–Jun. 2009.

[14] P. Lukowicz, "Wearable computing and artificial intelligence for healthcare applications." *Artif. Intell. Med.*, vol. 42, no. 2, pp. 95–98, 2008.

[15] P. Bonato, "Clinical applications of wearable technology," in *Proc. IEEE Int. Conf. Eng. Med. Biol. Soc.*, 2009, pp. 6580–6583.

[16] M. E. Pollack, "The use of AI to assist elders with cognitive impairment for an aging population," *AI Mag.*, vol. 26, no. 2, pp. 9–24, 2005.

[17] T. L. Westeyn, G. D. Abowd, T. E. Starner, J. M. Johnson, P. W. Presti, and K. A. Weaver, "Monitoring children's developmental progress using augmented toys and activity recognition," *Pers. Ubiquitous Comput.*, vol. 16, no. 2, pp. 169–191, Feb. 2012.

[18] T. de Jongh, I. Gurol-Urganci, V. Vodepivec-Jamsek, J. Car, and R. Atun, "Mobile phone messaging for facilitating self-management of long-term illnesses," *Cochrane Database Syst. Rev.*, vol. 12, no. CD007459, 2012.

[19] T.-J. Yun, H. Y. Jeong, T. D. Hill, B. Lesnick, R. Brown, G. D. Abowd, and R. I. Arriaga, "Using SMS to provide continuous assessment and improve health outcomes for children with asthma," in *Proc. 2nd Int. Health Informat. Symp.*, 2012, pp. 621–630.

[20] M. N. Burns, M. Begale, J. Duffecy, D. Gergle, C. J. Karr, E. Giangrande, and D. C. Mohr, "Harnessing context sensing to develop a mobile intervention for depression," *J. Med. Internet Res.*, vol. 13, no. 3, p. e55, Aug. 2011.

[21] Oxtext. (2014). True colours-improved management for people with biplar disorder. [Online]. Available: http://oxtext.psych.ox.ac.uk/

[22] Optimism. (2014, Jul.). Optimism apps. [Online]. Available: www.findingoptimism.com

[23] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong, "Moodscope: Building a mood sensor from smartphone usage patterns," in *Proc. 11th Annu. Int. Conf. Mobile Syst., Appl. Serv.*, 2013, pp. 389–402.

[24] A. Gruenerbl, G. Bahle, P. Lukowicz, and F. Hanser, "Using indoor location to assess the state of dementia patients: Results and experience report from a long term, real world study," in *Proc. 7th Int. Conf. Intell. Environ.*, 2011, pp. 32–39.

[25] A. Gruenerbl, G. Bahle, F. Hanser, and P. Lukowicz, "UWB indoor location for monitoring dementia patients: The challenges and perception of a real-life deployment," *Int. J. Ambient Comput. Intell.*, vol. 5, no. 4, pp. 45–59, Oct. 2013.

[26] T. Massey, G. Marfia, M. Potkonjak, and M. Sarrafzadeh, "Experimental analysis of a mobile health system for mood disorders," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 241–247, Mar. 2010.

[27] M. Frost, A. Doryab, M. Faurholt-Jepsen, L. V. Kessing, and J. E. Bardram, "Supporting disease insight through data analysis: Refinements of the monarca self-assessment system," in *Proc. UbiComp*, 2013, pp. 133–142.

[28] D. Tacconi, O. Mayora, P. Lukowicz, B. Arnrich, C. Kappeler-Setz, G. Tröster, and C. Haring, "Activity and emotion recognition to support early diagnosis of psychiatric diseases," in *Proc. 2nd Int. Conf. Pervasive Comput. Technol. Healthcare*, 2008, pp. 100–102.
[29] A. Gruenerbl, G. Bahle, J. Weppner, P. Oleksy, C. Haring, and P. Lukowicz, "Towards smart phone based monitoring of bipolar disorder," in *Proc. 2nd ACM Workshop Mobile Syst., Appl. Serv. HealthCare*, Nov. 2012, pp. 3-1–3-6.
[30] V. Osmani, A. Maxhuni, A. Gruenerbl, P. Lukowicz, C. Haring, and O. Mayora, "Monitoring activity of patients with bipolar disorder using smart phones," presented at the Int. Conf. Adv. Mobile Comput. Multimedia, New York, NY, USA, Dec. 2013.
[31] A. Gruenerbl, V. Osmani, G. Bahle, J. C. Carrasco, S. Oehler, O. Mayora, C. Haring, and Lukowicz, "Using smartphone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients," presented at the 5th Augmented Human Int. Conf., New York, NY, USA, Mar. 2014.
[32] A. Muaremi, F. Gravenhorst, A. Gruenerbl, B. Anrich, and G. Troester, "Assessing bipolar episodes using speech cues derived from phone calls," presented at the 4th Int. Symp. Pervasive Comput. Paradigms Mental Health (MindCare), Tokyo, Japan, 2014.
[33] S. Feese, A. Muaremi, B. Arnrich, G. Tröster, B. Meyer, and K. Jonas, "Discriminating individually considerate and authoritarian leaders by speech activity cues," in *Proc. Int. Workshop Social Behav. Anal. Behav. Change*, 2011, pp. 1460–1465.
[34] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile–The munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. Multimedia*, 2010, pp. 1459–1462.
[35] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 7, pp. 179–188, 1936.
[36] University of Waikato, "Data mining with open source machine learning software." *WEKA Mannual Version 3.6.10*, 2002–2013 (Online). Available: http://www.cs.waikato.ac.nz/ml/weka/

Agnes Grünerbl received the Dipl. Ing. degree in biomedical informatics from the University for Health Science and Technology, Tyrol, Austria, where she graduated in 2007.

She is currently working with the Embedded Intelligence group of Prof. Lukowicz at Technische Universität, Kaiserslautern, Germany, and with the DFKI where she is working toward the Ph.D. degree. Her main research interests include exploring new possibilities to assist health care and diagnosis with modern technology. As a result, she has already successfully deployed several real-life studies in health care. The technology used ranges from smartphone and location systems to smart carpets.

Amir Muaremi received the B.Sc. degree in electrical engineering from HSR Rapperswil, St. Gallen, Switzerland, in 2008, and the M.Sc. degree from ETH Zurich, Zürich, Switzerland, in 2010. During his studies, he completed the Master Thesis in the Speech and Audio Processing at Imperial College London, London, U.K. Since 2011, he has been working toward the Ph.D. degree in the electronics laboratory with ETH Zurich.

His research interests include speech and biomedical signal processing, and machine learning and pattern recognition.

Venet Osmani received the Ph.D. degree from TSSG, Waterford Institute of Technology. Ireland.

He is a Senior Researcher at CREATE-NET and a Lecturer at the University of Trento, Trento, Italy. His main research interests include mining of human behaviour data and use of this information in healthcare applications, specifically correlation of behaviour data with the disease. He has been a Guest Editor for a number of Special Issues in this area, has organized UbiHealth Workshop in Ubicomp Conference for a number of years, and is a Steering Committee Member of the Pervasive Health Conference.

Gernot Bahle received the Dipl. Inf. degree in computer science from the University of Passau, Passau, Germany.

After graduating with a Diploma in computer science as his major subject and minors in medicine and psychology, he was with Advanced Cad, Regensburg, Germany. Since 2009, he has been a part of the research group of Dr. P. Lukowicz, currently located at the German Research Center for Artificial Intelligence (DFKI). His research interests include context and activity recognition, abstract and formal context modeling, sensor fusion as well as learning and emergence in complex, and distributed and collaborating systems.

Stefan Öhler received the Graduate degree in psychology with an additional specification in sport psychology from the Leopol-Franzens-Universität, Innsbruck, Austria. He has been working with athletes in team sports and children displaying behavioral problems. Furthermore, he has worked on several international studies about suicide prevention in young people, education in gambling addiction, and dealing with bipolar disorder.

Gerhard Tröster studied electrical engineering in Darmstadt and Karlsruhe, Germany, and received the Doctorate degree in the design of analog integrated circuits from the Technical University of Darmstadt, Darmstadt, in 1984.

For eight years, he was at Telefunken (Atmel) Heilbronn, where he headed various national and international research projects centered on the key components for ISDN and digital mobile phones. Since 1993, he has been a Full Professor with the ETH Zürich, Zürich, Switzerland, and directs the Electronics Laboratory. In 1997, he cofounded the spin-off u-blox AG. His research interests include wearable computing for healthcare, wireless sensor networks, and smart textiles applying flexible and organic electronics.

Oscar Mayora received the Ph.D. degree at DIBE (Departimento di Ingeg neria Biofisica ed Electronical, University of Genoa, Italy).

He is the Head of Mobile and Ubiquitous Technologies Group in CREATE-NET.

Dr. Mayora is a Senior Member of the ACM and SIG-CHI and was a Former President of ACM SIG-CHI in Mexico. He is the Founder and a Permanent Member of the Steering Committee of the Pervasive Health Conference. He has participated as Scientific Project Coordinator of the MONARCA project on personal health systems for bipolar disorder treatment and is currently coordinating the NYMPHA-MD project on precommercial procurement on mobile applications and services for bipolar disorder treatment.

Chrisitan Haring is a Doctor of medicine, Psychiatrist, and Psychotherapist, and active in mental health care.

He is currently the Medical Head of the Department of Psychiatry and Psychotherapy with the State Hospital in Hall in Tirol, Austria. Besides being a Leading Member of several psychiatric and prevention institutions and committees, he is an Editor and Member of the board of numerous considerable professional journals. Furthermore, he supports ongoing research in psychiatry and has been a part of several international research projects in the recent years.

Paul Lukowicz received the M.Sc. (Dipl. Inf.) and Ph.D. (Dr. rer nat.) degrees in computer science and the MSc. degree in physics (Dipl. Phys.) both from the University of Karlrruhe Germany.

He was a Full Professor (W3) at the University of Passau, Passau, Germany. He is currently a Full Professor of artificial intelligence with the German Research Center for Artificial Intelligence and the Technical University of Kaiserslautern, Germany, where he heads the Embedded Intelligence group. His research interests include context aware ubiquitous and wearable systems including sensing, pattern recognition, system architectures, models of large-scale self-organized systems, and social interactive systems and applications.