A Survey on Mobile Social Signal Processing

NIKLAS PALAGHIAS, SEYED AMIR HOSEINITABATABAEI, and MICHELE NATI,
University of Surrey
ALEXANDER GLUHAK, Intel Labs Europe
KLAUS MOESSNER, University of Surrey

Understanding human behavior in an automatic but nonintrusive manner is an important area for various applications. This requires the collaboration of information technology with human sciences to transfer existing knowledge of human behavior into self-acting tools. These tools will reduce human error that is introduced by current obtrusive methods such as questionnaires. To achieve unobtrusiveness, we focus on exploiting the pervasive and ubiquitous character of mobile devices.

In this article, a survey of existing techniques for extracting social behavior through mobile devices is provided. Initially, we expose the terminology used in the area and introduce a concrete architecture for social signal processing applications on mobile phones, constituted by sensing, social interaction detection, behavioral cues extraction, social signal inference, and social behavior understanding. Furthermore, we present state-of-the-art techniques applied to each stage of the process. Finally, potential applications are shown while arguing about the main challenges of the area.

Categories and Subject Descriptors: A.2 [General and Reference]: Document Types; Surveys and Overviews; Human-centered computing; H.5.3 [Collaborative and Social Computing]: Ubiquitous and Mobile Computing

General Terms: Design, Theory, Human Factors, Performance

Additional Key Words and Phrases: Social signal processing, mobile phones, social behavior

ACM Reference Format:
Niklas Palaghias, Seyed Amir Hoseinitabatabaei, Michele Nati, Alexander Gluhak, and Klaus Moessner. 2016. A survey on mobile social signal processing. ACM Comput. Surv. 48, 4, Article 57 (March 2016), 52 pages.
DOI: http://dx.doi.org/10.1145/2893487

1. INTRODUCTION

Human behavior understanding has received a great deal of interest since the beginning of the previous century. People initially conducted research on the way animals behave when they are surrounded by creatures of the same species. Acquiring basic underlying knowledge of animal relations led to extending this information to humans in order to understand social behavior, social relations, and so forth. Initial experiments were conducted by empirically observing people and retrieving feedback from them.
These methods gave rise to well-established psychological approaches for understanding human behavior, such as surveys, questionnaires, camera recordings, and human observers. Nevertheless, these methods introduce several limitations including various sources of error. Completing surveys and questionnaires induces partiality, unconcern, and so forth [Groves 2004]; human error [Reason 1990]; and additional restrictions in the scalability of the experiments. Accumulating these research problems leads to a common challenge, the lack of automation in an unobtrusive manner.

An area that has focused on detecting social behavior automatically and has received a great amount of attention is Social Signal Processing (SSP). The main target of the field is to model, analyze, and synthesize human behavior with limited user intervention. To achieve these targets, researchers presented three key terms that constitute different levels of abstraction in the process of educing social behavior [Brunet et al. 2012; Vinciarelli et al. 2012a; Poggi et al. 2012]. Behavioral cues include various characteristics of human behavior that are extracted from a modality such as prosody of the voice and interlocutors’ spatial arrangement. The combination of these behavioral characteristics indicates a person’s current sentiment, understanding, attention, interest, and so forth, which are social signals. Pentland [2008] described social signals as nonverbal communication signals emitted when people are socially interacting. Merging these social signals in a longer temporal term leads to a person’s social behavior. In recent literature, the terms have been used in other areas such as social networks [Iamnitchi et al. 2012] to indicate every social-related Internet activity of a user. However, we do not consider this aspect in behavior inference. Social networks may function as an enhancement of SSP to provide additional information regarding the context, but in our interpretation the two areas are distinct.

In Vinciarelli et al. [2009], a generic procedure was proposed to detect social behavior:

1. Data capture
2. Person detection
3. Extraction of audio and visual behavioral cues, and their mapping to social signals
4. Incorporation of context to detect social behavior from social signals

This procedure is focused on detecting the social behavior of people through audio and visual data, from an external observer’s point of view. In order to achieve this observation, microphones and cameras are required to be deployed on the scene to monitor people. The major disadvantages of this approach are (1) the system’s intrusiveness as people feel monitored by seeing deployed obtrusive hardware such as microphones and cameras used primarily for surveillance in contrast to the utilization of pervasive devices such as smartphones; (2) limited mobility of the system, where in case of the requirement for conducting an experiment in a different area there is a need for redeploying and reconfiguring the system to the specific environment; (3) the confinement in scalability because the equipment is deployed at a certain environment and cannot follow the user’s mobility; (4) in audio and visual data, there is a need to perform person detection, while mining social interactions during which social signals are emitted is neglected; and finally (5) establishing ground truth in audio and visual data requires labeling, which is a time-consuming process and may induce human error.

1.1. A Mobile and Opportunistic Point of View

The purpose of this article is to review state-of-the-art techniques for extracting social behavior through mobile phones and also to introduce a discussion on the remaining challenges, existing gaps, and potential extensions of existing solutions of the area. Understanding social behavior in an automatic, nonintrusive, mobile, but also scalable manner constitutes a significant challenge with several potential applications. To address this challenge, close collaboration is required from the fields that accord two of the
most important components of the field, information technology and psychology. This collaboration will support the development of opportunistic nonintrusive self-acting tools for extracting human behavior. These tools will expunge several sources of error introduced by current obtrusive and user-engaging methods that incorporate human factors in the sensing process. In parallel, SSP focused on providing concrete solutions regarding modeling, analysis, and synthesis of social behavior. However, as mentioned, some major gaps have been identified.

In order to fill these gaps, we determined the following objectives, which will drive research on extracting personalized social behavior a step further:

—Utilizing nonintrusive approaches
—Capturing cues from a user’s perspective, to produce personalized data
—Leveraging multiple modalities, to extract more robust and reliable behavioral information
—Utilizing a continuous sensing and inference process, without mobility and scalability restrictions
—Eliminating the external hardware requirement

Smartphones have become a core feature of our daily lives. In recent years, the popularity and computational power of mobile phones have led to a new era where they are substituting computers and other means of communication such as old feature phones, fixed line phones, and so forth. Moreover, to facilitate a more rich user experience, mobile phone manufacturers have integrated various sensors such as an accelerator, gyroscope, GPS, digital compass, microphone, and camera. Furthermore, online application stores have given the opportunity to third-party developers to implement their own applications utilizing available integrated sensors seamlessly. Combining embedded sensors and application stores will introduce radical changes in fields such as healthcare, environment monitoring, and human behavior recognition by allowing easy, nonintrusive, and wide deployment of mobile applications.

Given the pervasive and ubiquitous character of mobile devices and considering the built-in sensing features, we consider smartphones as ideal devices for extracting social behavior among people. To support this claim, we introduce Mobile Social Signal Processing (Mobile SSP) while proposing the main architecture of human behavior inference for mobile applications. Further, each stage is analyzed by providing state-of-the-art techniques capable of being executed on mobile devices. Also, potential application cases will help to familiarize the reader with areas that will benefit from the growth of Mobile SSP, followed by a discussion of research opportunities that may be leveraged for further contribution to the field.

In the remainder of this article, a survey for Mobile SSP is provided. Section 2 describes the overall area of Mobile SSP while clarifying the core terms of the field. A brief description of existing sensing frameworks is shown in Section 3 and assists the reader in the selection criteria. State-of-the-art techniques utilized to detect social interactions among people on mobile phones are presented in Section 4. Behavioral cues extracted on smartphones while informing about their advantages and disadvantages are described in Section 5. Section 6 showcases methods of mining social signals and mapping them to social behaviors. Section 7 describes existing and potential applications of Mobile SSP. An overall discussion about methods presented for extracting social behavior on mobile phones is argued in Section 8. Finally, challenges of the area are outlined in Section 9, and the article concludes with Section 10.

2. AN OVERVIEW OF MOBILE SOCIAL SIGNAL PROCESSING

Mobile devices and in particular smartphones are ubiquitous. Multimodal sensing capabilities combined with increased computational power and available tools for mobile application development led to the view that smartphones are ideal devices for filling
the gap of lack of automation in social behavior understanding. Users can easily install an app from online application stores without any geographical restrictions and the device will automatically become a human-behavior-aware smartphone. To discriminate and categorize more easily the types of applications, two classes have been defined: participatory, which are the social behavior detection applications that require the user’s participation in the sensing process, and opportunistic, where the user is not involved in the process [Lane et al. 2010]. In order to minimize the obtrusiveness of the system and secure the user’s spontaneous behavior, the main attention of the work is focused on opportunistic social behavior detection applications.

In the area of SSP, users are continuously monitored through surveillance hardware; however, explicit cameras and microphones create the perception that the user is being monitored. Depending on the intrusiveness of the hardware utilized in the sensing process, people’s spontaneous behavior may be less affected, which may allow them to adapt faster. An inherit characteristic of the area of Mobile SSP is that users are being monitored but with pervasive devices such as smartphones. The main usage of smartphones relates to everyday routine processes such as phone calls, messaging, web browsing, and gaming. As such, monitoring through these types of devices may affect less the spontaneity of users’ social behavior contrary to the utilization of dedicated surveillance hardware. Also, as users install the social behavior app, they give their permission for each type of sensed data, as opposed to the case of surveillance hardware, where monitored users do not have any control over the sensed data. Nevertheless, independent of the sensing technology, the spontaneity of people feeling monitored may be affected. Taking this fact into account is crucial for the understanding of social behavior [Pentland 2008].

Similar to Vinciarelli et al. [2010] and based on the literature review, we concluded that the following steps need to be taken for extracting social behavior on mobile devices (see Figure 1).

1. Sensing
2. Social interaction detection
3. Extraction of behavioral cues
4. Understanding social behavior by inferring social signals

Social behavior inference on mobile devices is initiated by the sensing process. During daily life, users emit behavioral cues and social signals, which are captured by sensors of the mobile device. These sensors may be integrated into the device or enclosed in
external hardware that communicates with the mobile device. Each sensor detects a particular modality, and then it converts the detected signal into a raw data signal. The result is processed into a desired format or is directly forwarded to the next stage of social behavior inference. Researchers have developed various sensing frameworks to allow developers to collect data in an abstract and uniform way, while in some cases they also include an inference engine.

Pentland recognized the emission of social signals during a social interaction [Pentland 2007]. This signifies the importance of recognizing social interactions before initiating the process of social behavior understanding. After retrieving the appropriate data from the mobile device’s sensors, social interaction detection may be performed as a preprocessing step of social behavior inference. Understanding social interactions provides important contextual information that may be leveraged in the next steps of social behavior inference. The knowledge of ongoing social interactions may also be utilized for filtering data and allowing the development of adaptive sensing and inference techniques. In applications focusing on extracting behavioral information not related to the social aspect of a person, it is strongly encouraged to include this step as it provides important contextual information.

Following the identification of ongoing social interactions is the extraction of behavioral cues. Different modalities may be leveraged for the extraction of a behavioral cue, depending on the grammar defined in psychology. Each selected sensed modality is forwarded to behavioral cues extraction. Existing literature has been classified into seven categories based on the types of cues each work extracts (see Figure 1). The behavioral cues extraction is achieved through techniques such as decision models, statistical analysis, and so forth.

The final stage of Mobile SSP is the transition from the understanding of social signals to social behavior inference. Close collaboration with social sciences may provide the theoretical mapping among behavioral cues, social signals, and social behaviors. Literature has been grouped based on the inferred social behavior through mobile phones. The extracted behavioral cues are fed into decision-making techniques to mine social signals and infer long-term social behavior.

To facilitate the reader’s understanding of the field, we provide an outline of the main steps and requirements for an integrated and real-world-enabled Mobile SSP:

—Define the context of the Mobile SSP application.
—Select the modalities required to infer a particular social behavior.
—Define the grammar of behavioral cues and social signals that will lead to social behavior inference.
—Evaluate and verify the reliability of the approach in a real-world environment based on ground truth.

In addition to these requirements, researchers need to consider the intrusiveness, security, and privacy of the system. Researchers need to take into account the computational burden and energy consumption that may endanger user experience. These parameters do not constitute a prerequisite for the realization of Mobile SSP but will facilitate the user experience and privacy.

In the following sections, each of the predefined stages will be analyzed and state-of-the-art research outlined. The works described in the next sections are summarized in the Online Appendix, introducing the techniques developed in each stage of social behavior inference.

3. SENSING FRAMEWORKS

Sensing is the first stage in extracting human behavior on mobile devices. In this stage, selection of appropriate modalities is performed. These will later on be processed and
analyzed to reveal information about users’ social behavior. It constitutes the lowest level of the process, which collects raw data from sensors and other interfaces that can provide information relevant to the user (see Figure 2). After retrieving information from sensors, either the raw data are forwarded to the next stages or lightweight and simplistic processing may be performed to minimize the complexity and computational burden at the upcoming stages. As shown in Figure 1, the next stages in social behavior inference may be performed either on the device or at a back-end server.

This section introduces and then compares existing sensing frameworks. Through this introduction, the reader should be able to understand the criteria based on which the sensing framework should be selected for a desirable social behavior application. An extensive analysis of existing sensing frameworks is outside the scope of this article and the reader is referred to Lane et al. [2010] and Hoseinitabatabaei et al. [2013].

3.1. Data Collection

This subsection focuses on sensing frameworks that perform only data collection and some minor preprocessing allowing the extraction of human behavior offline. An application is deployed on the device, which collects data from preconfigured sensors and then either stores the information on the device’s internal storage or performs uploading to a back-end server. The retrieved data are merged and forwarded to the next stage in order to extract behavioral cues. Table I summarizes existing frameworks for data collection by presenting (1) the type of sensor data retrieved, (2) the incorporation of a mechanism for energy efficiency such as adaptive sensing, (3) the embodiment of a privacy-preserving approach, and finally (4) the software license through which the authors released a particular framework.

*MyExperience* [Froehlich et al. 2007] retrieves and logs contextual information from physical and virtual (e.g., phone usage patterns) sensors. *SeeMon* [Kang et al. 2008] investigates the context of the device and adapts the sensing process by mining local sensors and installed applications in an energy-efficient manner. To preserve privacy in crowd-sensing applications, Cornelius et al. [2008] introduced *Anonymsense*, which permitted, through a centralized sensing coordination, the collection and storage of anonymous smartphone sensed data in a collaborative manner. *OpenDataKit* [Anokwa et al. 2009] is a set of tools for sensing and aggregating data from mobile phones.
Furthermore, Das et al. [2010] developed PRISM, a platform for dynamic deployment of participatory and opportunistic sensing applications on mobile phones while maintaining security through a sandboxed environment. Another approach that focuses on privacy preservation but also on energy efficiency for long-range-duration deployment is LiveLab [Shepard et al. 2011], which stores inertial, phone usage, positioning, and network-based data. SystemSens [Falaki et al. 2011] is a data retrieval tool that enables researchers to capture mobile data in large-scale experiments. Funf [Funf 2011] is a mobile data collection platform developed by MIT Media Labs, procuring easily configurable sensing and data retrieval capabilities. Medusa [Ra et al. 2012] achieved crowdsensing and simultaneous coordination of multiple mobile devices. METIS [Rachuri et al. 2013] is a distributed system that decides, based on the device status and user context, to perform on-device or infrastructure-oriented sensing. MSF [Cardone et al. 2013] is a recent data collection framework that complies with multipipeline architecture and targets in providing an abstraction regarding the sensing process.

**Discussion.** Regarding the data collection frameworks, MyExperience [Froehlich et al. 2007] constitutes an event-triggered approach that is energy efficient and does not require any polling process to identify state changes but does not utilize data from inertial sensors. OpenDataKit [Anokwa et al. 2009], SystemSens [Falaki et al. 2011], and Funf [Funf 2011] are three distinct configurable open-source data collection tools that allow offline merging of data from different sensors. However, they do not perform adaptive sensing based on the context to improve energy efficiency. This is critical for these types of applications. Medusa [Ra et al. 2012] allows a coordinator to retrieve a certain type of sensor data from a specific device. Furthermore, METIS [Rachuri et al. 2013] is the first work that lightens a mobile device by selectively performing sensing through the infrastructure but simultaneously narrows the mobility and increases the intrusiveness of the system. MSF [Cardone et al. 2013] focuses on easing the development of sensing applications. It requires the designer to implement the sensing functionality but handles tasks, power management, and resource allocation. However, it does not constitute a ready-to-deploy solution and is suitable only for developers.

### 3.2. Inference Engines

As opposed to the previous subsection, this section includes frameworks that perform sensing and inference on the device or on a back-end server. The sensor selection is predefined or configurable depending on the implementation. Data are retrieved from the sensors and forwarded to an inference pipeline. Based on the inference, one or more preprocessing stages could be performed. Then, the appropriate intelligence is
Table II. Inference Engine Frameworks for Online Analysis

| Framework               | Detection          | Energy Efficient | License   |
|-------------------------|--------------------|------------------|-----------|
|                         | Activity | Voice | Location | Emotion | Sociability |
| CenceMe [Miluzzo et al. 2008] | ✓        | ✓     |          |         |             |
| BeTelGeuse [Kukkonen et al. 2009] | ✓        |       | ✓        |         |             |
| Jigsaw [Lu et al. 2010] | ✓        | ✓     | ✓        |         | ✓           |
| EmotionSense [Rachuri et al. 2010] | ✓        | ✓     | ✓        | ✓       | ✓           |
| SociableSense [Rachuri et al. 2011] | ✓        | ✓     | ✓        | ✓       | ✓           |
| AmbientDynamix [Carlson and Schrader 2012] | ✓        | ✓     | ✓        |         |             |
| Auditeur [Nirjon et al. 2013] |          | ✓     |          |         |             |

applied to retrieve the requested knowledge. An optional postprocessing phase, such as the consideration of historical inferences, outlier detection, or smoothing, may be applied to remove results that deviate from normal. Table II presents the state-of-the-art inference engines and identifies (1) the type of information that is extracted by each framework, (2) the development of an energy-efficient approach, and (3) the software license of the framework.

CenceMe [Miluzzo et al. 2008] is a distributed platform that performs multimodal sensing through mobile phones. A classification-based technique decides about inferring social context on the device or on a back-end server. It also allows the user to publish the inference to social networks. BeTelGeuse [Kukkonen et al. 2009] was one of the first tools that had the native capability of sensing and inferring about users’ context. Jigsaw [Lu et al. 2010] is a mobile platform that allows continuous data collection in an energy-efficient way, through multiple pipelines (one for each modality) and adaptive sampling based on user behavioral patterns. EmotionSense [Rachuri et al. 2010] is a framework for inferring user emotion and incorporates an intelligent engine for adapting the sensing process. As an extension of it, SociableSense [Rachuri et al. 2011] measures the sociability of people and introduces an adaptive inference mechanism (locally or distributed) based on reinforcement learning. AmbientDynamix [Carlson and Schrader 2012] is an equally important framework that allows the deployment of custom inference modules in a sandboxed environment. Also, Auditeur [Nirjon et al. 2013] is a context recognition framework that is focused only on the audio perspective, but it provides a collection of inference mechanisms for the specific modality.

Discussion. Regarding state-of-the-art sensing frameworks that have a human behavior inference, CenceMe [Miluzzo et al. 2008] performs preliminary detection of activity and conversation. However, only an application that publishes user context to social networks is publicly available. BeTelGeuse [Kukkonen et al. 2009] focused mainly on the sensing process, enabling the integration of external Bluetooth-connected sensors. It also allowed the incorporation of inference through plug-ins while initially providing location and activity classification components. Jigsaw [Lu et al. 2010] limits its sensing capabilities to the accelerometer, microphone, and GPS but provides integrated...
classification techniques for activity and voice recognition. Through a multithreaded approach, they try to limit the computational burden on the device due to the classification process. **EmotionSense** [Rachuri et al. 2010] and **SociableSense** [Rachuri et al. 2011] are based on the same framework, providing a quantification method for the user's emotion and sociability while performing adaptive inference through learning techniques. It is available for developers but also directly for less technical people.

Furthermore, **AmbientDynamix** [Carlson and Schrader 2012] allows the user to select existing or concrete components, integrate them in a main skeleton application, and perform the desirable social behavior detection. If the component exists, it constitutes an easy and reliable solution, while if the module requires development, it can be contributed to the community for further reuse. All the processes of sensing configuration, data logging, and resource management and concurrent procedures are handled in a seamless manner by the skeleton application, which reduces the developer's effort. If the targeting system is focused on mining social behavior information through audio data, **Auditeur** [Nirjon et al. 2013] constitutes a reasonable solution that provides the appropriate mechanisms to extract audio features but also allows the configuration of the desired classifier. In addition, it includes state-of-the-art techniques for contextual sound recognition.

### 3.3. Framework Comparison

A notable amount of works targeting sensing frameworks for mobile phones was briefly described in the two previous subsections. The literature was classified based on whether the framework enabled human behavior inference or not.

Overall, the first step in the design of a social behavior detection application is the decision about the sensing framework. Many researchers start by designing and developing the sensing process from scratch. However, as shown, works on sensing frameworks have reached a certain maturity, which allows component reuse. These frameworks provide off-the-shelf solutions for resource management, concurrency, data handling, energy efficiency, and concrete structure of the application. This should be leveraged in order to reduce the development time cycle and human error and increase code reuse. Most of them are released with open licences, allowing for clear understanding and editing but also contributing of the source code from the research community. Selecting a data collection or inference engine framework is highly dependent on the targeting application and how sufficient the capabilities are of each framework with respect to the researchers' envisioned outcome. Thus, it should be noted that selecting a certain framework does not lead to a right or wrong decision but results in a tool that will provide more or less enabling capabilities for developing a social behavior detection application.

### 4. SOCIAL INTERACTION DETECTION

The next stage of retrieving data from sensing the context is recognizing ongoing social interactions. People are assumed to interact socially when they are in close interpersonal distance, facing each other, and participating in a conversation. Pentland's definition of social signals [Pentland 2007] is that they are nonverbal communication signals that are conveyed when people are socially interacting. Thus, identifying possible social interactions accurately is an important stage of social behavior understanding and requires tackling.

---

1The classification of the techniques between the sections of social interaction detection (Section 4) and the behavioral cue of interpersonal distance (Section 5.5.1) has been performed based on the context in which they are utilized in the literature. It should be noted that the social interaction detection techniques that are based only on proximity sensing may also be utilized for extracting the behavioral cue of interpersonal distance and vice versa.
Researchers have developed several techniques to detect social interactions through smartphones. These techniques vary based on the level of accuracy and modalities included in the inference process. Among these approaches are single modality, which include coarse- or fine-grained distance estimation through Bluetooth and WiFi interfaces, audio-based distance, or relative position estimation. There is also multimodality, where different modalities are combined in the inference process to indicate if people are interacting (see Figure 3).

4.1. Single Modality

The majority of commercial off-the-shelf (COTS) smartphones today contain wireless communication interfaces such as Bluetooth and WiFi. Due to their wide availability, researchers often utilize them in order to detect when people are in proximity. By acquiring information about people’s proximity to each other, researchers made a strong assumption of interpreting the knowledge of proximity into the existence of social interaction. An implicit assumption is that every person is required to carry a smartphone. For the determination of proximity, several techniques have been proposed.

The most common approach is to perform discovery through one of the two interfaces, log the Bluetooth IDs (BTIDs) or WiFi Service Set Identifier (SSID), and classify all the detected nearby devices as social interactions. This method was applied in miscellaneous works to estimate when people are interacting with each other. Some examples of these works are Serendipity [Eagle and Pentland 2005], CenceMe [Miluzzo et al. 2008], and SoundSense [Lu et al. 2009]. Antoniou et al. [2011] aggregated the discovered smartphones based on BTIDs with static nodes. The Bluetooth ranges were overlapping to improve social interaction detection and provide some information about coarse-grained localization. The accuracy of this method is limited to the range of the communication means; that is, for Bluetooth, the minimum nominal range is around 10m [Kotanen et al. 2003], and for WiFi, the typical range is approximately 35m for indoor environments. Thus, every device/person detected is classified as being in a social interaction. It should be noted that these works do not provide error analysis of this social interaction detection approach.

---

Footnote:

2 Bluetooth and WiFi ranges are highly dependent on the surrounding environment and chipsets’ characteristics of devices performing the discovery and the detected devices.
The previous method introduces a noticeable amount of error. For that reason, researchers focused on developing distance estimation techniques that would remove a percentage of error from the previous approach by limiting the communication mean range. So, for detecting social interactions, Osmani et al. [2014] developed a machine-learning-based technique to estimate if users were in proximity or not by retrieving the WiFi Received Signal Strength Indicator (RSSI). They trained a model based on the maximum and mean value of a 20-sample window of WiFi RSSI achieving a median error of 0.5m for social interaction detection. In particular, they detected ongoing social interactions with 86% accuracy and true negatives with 84% accuracy.

Another approach for detecting social interactions through Bluetooth RSSI-based distance estimation was presented in Liu et al. [2013a]. They developed a probabilistic model for indoor and outdoor environments. It utilized some predefined empirically calculated thresholds to compute the probability of being in proximity to socially interact, with a claimed error rate of 4.3%. In detail, they retrieved an RSSI measurement that was smoothed through an exponential window moving average (EWMA) and a smoothing factor of 0.5. To compute the thresholds, they retrieved Bluetooth RSSI measurement in different environments and distances. Bluetooth technology natively operates in a mode that allows the device to discover but also to be discoverable by other devices without any firmware modifications. A less complex approach was presented in Hu et al. [2013], namely, MAUC. This work focused on detecting social interactions through Bluetooth RSSI thresholds, showing a detection rate over 90%. As opposed to previous approaches, it incorporated an adaptive discovery interval scheme, dependent on the user’s activity movement.

Discussion. Scientists tackled the detection of social interactions on mobile phones with different granularities. The discovery of nearby devices (e.g., Serendipity [Eagle and Pentland 2005], CenceMe [Miluzzo et al. 2008], etc.) is easily implementable. However, it provides an increased number of false positives due to inaccuracy in the interpersonal distance estimation, unawareness of spatial arrangement, and conversation existence. As an example, Figure 3(a) shows four people in vicinity, where only (A) and (B) are interacting, although all four are in discovery range, and they are thus classified as socially interacting. The WiFi interface on COTS smartphones cannot operate simultaneously in discovering and discoverable mode, as opposed to WiFi Direct. For that reason, the authors [Carreras et al. 2012; Matic et al. 2012] were forced to modify the firmware of the devices to switch between the two modes. As an improvement, several works tried to estimate the interpersonal distance of users, to infer the existence of social interactions. For distance estimation based on RF (Bluetooth or WiFi) signals, RSSI has been leveraged in order to create empirical models, mainly due to simplicity in implementation. However, RSSI measurements received on mobile phones showcase great fluctuation, which is affected by different environments, obstacles, human body absorption, reflections, and so forth. Machine-learning techniques constitute a notable effort to tackle the RSSI fluctuation [Osmani et al. 2014]. Threshold-based approaches [Liu et al. 2013a; Hu et al. 2013] usually require adjustment of the algorithm's boundaries based on the device and the environment. Other techniques such as Time Difference of Arrival (TDoA), Angle of Arrival (AoA), and Time of Arrival (ToA) provide significant limitations such as firmware modification for device time synchronization, lack of multiple antennas, and the need for external hardware and therefore are not recommended for COTS mobile phones.

From a simple discovery-based method, researchers have improved the social interaction detection through proximity. However, the assumption that when people are in proximity they are definitely interacting is strong. Hence, there is an imperative need to add other modalities in the inference process that will add new parameters such as spatial arrangement and conversation existence detection.
4.2. Multimodality

To improve social interaction detection based on a single modality, researchers started to incorporate other modalities. These modalities enhanced social interaction detection by providing information about users’ relative orientations (i.e., if they are facing each other or not) and about the conversation existence (i.e., if the users are really having a conversation and they are not two people having a spatial formation suitable for social interaction) (see Figure 3(b)).

An important attempt to identify the relative spatial arrangement of the users combined with proximity detection was Virtual Compass [Banerjee et al. 2010]. This system utilizes multiple RF interfaces such as Bluetooth and WiFi in order to estimate the interpersonal distance among users but also create a 2D localization map based on users’ relative distances. For distance estimation given RSSI measurements from both interfaces, they computed the average distance and the uncertainty based on the difference of the 90th and 10th distance percentile. Then, the authors applied regression on these features for distance estimation and achieved an error margin of 1.41m. For relative map localization, they propose a quick technique to estimate the initial coordinates of each user and then they apply an iterative method [Dabek et al. 2004] to refine the initially reckoned coordinates.

Matic et al. [2012] argued about a slightly different multimodal approach that incorporated interpersonal distance estimation with relative orientation calculation and conversation detection. In this work, the authors endeavored to increase the accuracy of social interaction detection by taking into consideration users' facing directions. The knowledge of users’ facing directions with respect to the earth’s coordinates allowed them to compute the relative orientation of each pair of users, in order to understand if they had the appropriate spatial arrangement to interact. They estimated the interpersonal distance of the users through an initial calibration phase, which led to a proximity detection model. During that period, they collected WiFi RSSI measurements at 1m distance and then, based on an indoor path loss model (PLM), they created an artificial dataset for distances 0.5m, 1m, ..., 5m. Based on this dataset, they computed the mean and maximum value of a 20-sample window and trained a Naive Bayes with Kernel Density Estimator (KDE) model for proximity detection. An external accelerometer was attached to a user’s chest to monitor his speech activity by performing spectral analysis of the signal it produced.

Discussion. In order to achieve higher accuracy and robustness in various environments, researchers combined various modalities. Virtual Compass [Banerjee et al. 2010] combined WiFi and Bluetooth to improve distance estimation and provide relative spatial arrangement detection. However, the simultaneous utilization of these interfaces cannot be utilized for continuous sensing due to high energy consumption and lack of ad hoc communication in currently available WiFi on COTS mobile phones. For further improvement of social interaction detection on smartphones, Matic et al. [2012] designed a work that provides a relatively accurate approach. As Figure 3(b) indicates, this method is able to identify correctly that (A) and (B) are socially interacting while (C) and (D) are not. They tracked users’ orientation and logged speech activity through an accelerometer attached to users’ chests. Yet, the smartphone was in a fixed body position and the external accelerometer was intrusive as it was placed on users’ chests.

4.3. Apposition of Social Interaction Detection Approaches

In this section, detection methods for social interactions among people were surveyed by presenting different approaches that researchers have developed to tackle this problem. It is important to understand the strengths and the limitations of each method.
As mentioned, the single-modality works provide the benefit of low complexity and unobtrusiveness through limited constraints regarding the wearing position and the lack of external hardware. However, the approach is characterized by a large amount of false-positive errors, which, depending on the application, could be neglected or require tackling. In case this error is not acceptable, a multimodal approach is more suitable that incorporates the user's facing direction and conversation detection. Although these multiple modalities provide additional information to tackle the social interaction detection problem, researchers may consider the accumulated error introduced by each modality. Through efficient fusion of several modalities, the error of each individual modality may cancel each other out, driving the system to a less erroneous approach.

Despite the numerous works on attempting to detect social interactions on mobile phones, to our knowledge currently there is no tool that may be utilized in a real-world environment, without any constraints and with minimum intrusiveness. Ideally, the framework may not require any firmware modification. Also, it would be able to be deployed on off-the-shelf smartphones, providing realistic and robust accuracy in a real-world environment, depending on the target application given only the integrated sensors of the device. Finally, both approaches are characterized by a tradeoff among user friendliness, system complexity, and accuracy [Madan and Pentland 2006] that should be considered by researchers depending on the needs of each application.

5. BEHAVIORAL CUES EXTRACTION

Social interaction detection provides personalized information about who is interacting with whom. As Figure 4 shows, during a social interaction, interlocutors emit cues such as spatial arrangement, posture, and gestures, indicating social signals such as intimacy, interest, and mirroring. Acquiring this knowledge leads to the next stage of mining human behavior on mobile devices, which is the extraction of behavioral cues. In this process, data obtained by the sensing procedure are preprocessed or classified through a machine-learning technique to retrieve some features that will be utilized in the next stage, the inference of social signals.
5.1. Auditory

In SSP, the literature has focused mainly on extracting social signals from audio and visual data. This fact procures a reasonable indication regarding the importance of behavioral cues extracted from audio data. This section aims to provide a brief overview of the techniques applied on acoustic data recovered from ambient sensors and especially from a mobile device’s microphone. The techniques are categorized based on the type of features extracted. The same classification was performed in Vinciarelli et al. [2009]. Thus, following and extending that taxonomy for techniques applied on mobile devices was considered as a reasonable continuity. It is noted that social signals refer to nonverbal communication signals emitted when a person is socially interacting [Pentland 2007]. Hence, natural language processing is excluded from Mobile SSP as it is considered a separate field.

5.1.1. Prosody. This behavioral cue provides information about the characteristics of a person’s voice when socially interacting, that is, phrasing, stress, and intonation [Selkirk and Goldsmith 1995]. To infer about the prosody of a person’s voice on mobile devices, the literature has focused on extracting certain features. Razak and Abidin [2005] extracted prosodic features such as energy, linear predictive coding (LPC) coefficients, duration, pitch, and jitter of each recorded frame. VibeFones [Madan and Pentland 2006] focused on pitch, amplitude, mean energy, frequency of the fundamental format, and spectral entropy. AMMON [Chang et al. 2011] calculated the zero crossing rate (ZCR), root mean square, frame energy, pitch, harmonics-to-noise ratio, and Mel-Frequency Cepstral Co-efficients (MFCC). StressSense [Lu et al. 2012] reported that pitch and its derivatives were the most informative features, followed by jitter, spectral centroid, high-frequency ratio, speaking rate, and MFCCs. In SoundSense, Lu et al. [2009] extracted ZCRs, low-energy frame rates, and other spectral features that were fed in a J48 decision tree [Quinlan 1993]. The most recent work SocioPhone [Lee et al. 2013b] calculated prosodic features through pitch, energy, loudness, rhythm, and spectral features (formants,3 bandwidths, spectrum intensity).

Discussion. The optimal prosodic feature set varies based on the target application. Razak and Abidin [2005] evaluated different prosodic feature sets. Those that included LPC coefficients had the best performance, while the set with speech energy, duration, pitch, and jitter performed worst. The authors did not apply any feature reduction technique on the training data, based on various factors such as information gain to retrieve the most informative feature set [Hall and Holmes 2003]. VibeFones [Madan and Pentland 2006] requires a long-term analysis to derive standard deviations of the features and does not describe the process of concluding to the particular feature set. AMMON [Chang et al. 2011] showed the performance improvements achieved when combining prosodic features with glottal timings.4 However, the system was evaluated offline on datasets created in constrained environments, while the performance in real-world situations was not provided. StressSense [Lu et al. 2012] selected a particular feature set based on information gain. The authors did not provide any quantitative analysis of energy consumption of the system as they extracted cumbersome features including MFCC. It should be noted that although MFCC improved the accuracy of the system, these features allow rebuilding of speech segments and further natural language processing. This fact induces some privacy issues as opposed to features such as pitch, speaking rate, and jitter, which do not allow derivation of speech segments. SoundSense [Lu et al. 2009] focused on features that are not affected by the volume;

3Formant in a vocal signal is the accumulation of acoustic energy close to a certain frequency.
4Glottal timings refer to the air flow variations produced during speech.
in spectral features, they removed DC components. To preserve users’ privacy, they performed data processing on the device and then discarded any raw audio readings. *SocioPhone* [Lee et al. 2013b] is able to cope with ambient noise distributed uniformly to nearby devices but does not incorporate any on-body position detection mechanism of the device to discard positions, such as bags, that degrade the quality of the raw sound signal.

In the literature, many works inferred that features such as pitch and its statistics were considered as the most informative features. Additionally, features such as speaking rate, MFCCs, energy, and spectral characteristics were also included in the process of detecting the vocal prosody of a user. The spectral-based features constitute a common ground in various works and especially formants, bandwidths, and intensity. The disadvantage of these features is the requirement of transforming the time-based values into frequency-based values before every inference, which induces an additional computational burden. Only a specific set of the 20 MFCCs are mainly selected during the inference process depending on the classification target. Before including MFCCs in the feature set, a designer may consider the accumulative burden of these coefficients due to a computationally demanding extraction procedure.

5.1.2. Conversation Versus Silence. During a social interaction, speech and silence operate as regulators of a conversation emitting social signals such as consensus and rejection and reveal interlocutors’ social behavior including their emotions [Koudenburg et al. 2011]. One of the well-known and widely used techniques to infer conversation existence was presented by Basu [2003]. It specified a linked Hidden Markov Model (HMM) with three features: noninitial maximum of the normalized noisy autocorrelation, number of autocorrelation peaks, and normalized spectral entropy. The first layer of the model infers regarding voice existence and the second layer regarding speech occurrence. This technique was adopted by Choudhury and Basu [2004], *Vibefones* [Madan and Pentland 2006], *StressSense* [Lu et al. 2012], and *MeetingMediator* [Kim et al. 2008; Wyatt et al. 2011]. Another technique widely used by systems such as SpeakerSense [Lu et al. 2011] and *Auditeur* [Nirjon et al. 2013] is to calculate the ZCR of an audio frame and then apply a classification method to infer if the segment contains speech [Saunders 1996].

Matic et al. [2012] inducted a privacy-oriented approach that incorporated an accelerometer on a user’s chest. By analyzing the sum of the power spectral densities, computing the integral and mean of the frames (80–256Hz), and feeding it to Naïve Bayes with KDE [John and Langley 1995], they were able to detect when the user was speaking. CoenoFire [Feese et al. 2013] detects speech through *Long-Term Signal Variability* (LTSV) presented in Ghosh et al. [2011]. Also, *AutoSense* [Ertin et al. 2011] utilised a Respiratory Inductive Plethysmograph (RIP) in order to compute lung volume and breathing rate, from which they detected conversation existence.

**Discussion.** For detecting conversation existence or absence through a mobile device, research has focused on audio-based and accelerometer-based data. The conversation detection methods based on the microphone data funnel in Basu [2003] or through some preprocessing steps train a Gaussian Mixture Model (GMM) that identifies (non)conversation segments. Basu [2003] discusses a well-established approach for detecting speech in raw microphone data, which achieves less than 10% error estimations at 6.4m even with increased interpersonal distance. Nirjon et al. [2013] brought together the most popular preprocessing steps in order to train a model for conversation and silence detection. A postprocessing step may be applied to add time dependence through an HMM. Among the preprocessing steps utilized, performing Fast Fourier Transform (FFT) and extracting MFCC features constitute the most energy-consuming processes among the state-of-the-art techniques, as opposed to ZCR,
which is a simple and robust feature. Matic et al. [2012] and AutoSense [Ertin et al. 2011] induced two privacy-preserving approaches\(^5\) for speech recognition through accelerometer (93% accuracy) and RIP data (over 87% accuracy), both evaluated in real-world situations. Although Matic et al. [2012] achieved around a 10% error rate in real-world environments, it is prone to coughing and various means of transportation that confer vibrations. Additionally, attaching an external sensor on a user’s chest is considered intrusive. CoenoFire [Feese et al. 2013] focused on LTSV, which is suitable for noisy environments but is not able to discriminate speech among various users. Finally, classifying conversation existence is a process that can be applied on mobile phones as shown in the literature; nevertheless, including energy-consuming preprocessing steps will increase the computational burden.

5.1.3. Turn Taking and Vocal Outbursts. In linguistics, turn taking refers to the process of exchanging speech turns during a conversation, including speech overlap and showing the willingness of a person to continue a conversation [Schegloff 2000]. The occurrence of nonlinguistic vocalizations (i.e., vocal outbursts) provides additional information about interest, boredom, willingness to continue a conversation, and so forth among the interlocutors [Schrder 2003]. Given a set of features, turn-taking detection is identified mainly by training a GMM for each speaker through an Expectation Maximization (EM) algorithm, to allow classification of the most probable speaker for a certain speech frame. This process is called speaker recognition and is harnessed to understand between which people turn taking is occurring. Miluzzo et al. [2010] developed an online speaker diarization system that in a distributed manner infers interlocutors’ turn taking through the previously described typical speaker recognition pipeline. Aran and Gatica-Perez [2011] perform speaker diarization\(^6\) and then calculate three types of turn-taking features: (1) independent (speaking length, number of speaking turns, turn duration statistics), (2) relational (interruptions, order, centrality), and (3) meeting (number of silent moments, overlapped speech). A similar approach was followed by SocioPhone [Lee et al. 2013b]; instead of meeting features, it incorporated interaction features that included the duration of speaking and nonspeaking turns. For detecting vocal outbursts, VibePhones [Madan and Pentland 2006] considered the distribution of utterance length (i.e., z-score).

Discussion. Understanding turn taking in an audio data sequence requires the execution of speaker recognition and then identification of the segments with speaker change or overlapping. The features selected are the same utilized in conversation detection. A great deal of attention should be paid in selecting the optimal dataset with respect to the application context in order to achieve high accuracy and robustness in the inference process. A certain model is trained for each speaker, inducing the requirement of a speaker model library if the system extracts information from all the interlocutors. In the case of adding a new user to the system, a model has to be trained especially for this user and incorporated to the library. Then a Maximum Likelihood algorithm is applied to identify the most likely model. The speaker diarization, in most cases, is executed offline, where all audio segments have been logged and categorized to each speaker such as in Aran and Gatica-Perez [2011]. For on-line execution, the process requires a connection with the centralized library in order to identify the speakers and then log information about turn taking. SocioPhone [Lee et al. 2013b] performs turn-taking

\(^5\)Matic et al. [2012] utilize accelerometer data and AutoSense [Ertin et al. 2011] senses lung volume and breathing rate; thus, they are considered privacy-preserving approaches as they do not focus on audio data that allow natural language processing.

\(^6\)Speaker diarization refers to the process of speaker recognition followed by clustering with respect to each speaker.

ACM Computing Surveys, Vol. 48, No. 4, Article 57, Publication date: March 2016.
detection online; it achieves its highest accuracy when the number of devices is equal to the interlocutors and the devices are placed on a meeting table. When the number of mobile devices is reduced or the device is placed in a bag or in trousers pockets, the accuracy is degraded. In contrast, Miluzzo et al. [2010] were able to maintain similar accuracy even when the number of smartphones was reduced, but the recognition accuracy was degraded in short-term turn takings. Loading all speaker models on the device will increase the computational burden and energy consumption and will degrade the user experience. On the contrary, off-loading this computation burden to the cloud will induce a certain communication cost in terms of energy consumption and cost increase in users’ data plans. After performing speaker diarization, statistical analysis such as speaking time, number of turns, and so forth can provide valuable information about the sociability and the overall social behavior of the user. Regarding vocal outbursts, VibePhones [Madan and Pentland 2006] utilized the z-score of utterance length, which is an easy-to-extract feature but requires the knowledge of mean and deviation of the population to compute the particular feature.

5.1.4. Speech Activity. Another auditory cue is speech activity, which includes various derivatives of the time a person is talking in a conversation (i.e., accumulated speech time, speaking frequency, etc.), indicating social behaviors such as sociability and dominance. VibePhones [Madan and Pentland 2006] detects a user’s activity in voice, by performing initially speech recognition and then calculating the z-score of the time the user was speaking. In CoenoFire [Feese et al. 2013], having detected speech in audio data, the authors compute the speaking time for each user, which indicates the speech activity of each participant.

Discussion. Speech activity is a feature that can be derived from previous auditory cues. Literature has extracted this cue by initially detecting voice and speech through the aforementioned techniques. Further, by inferring the speech segments that belong to the user through a personalized speech detection model, an accumulation of the speech windows is performed through statistical analysis. This accumulation refers to the computation of speech frequency, overall speech time, and so forth. VibePhones [Madan and Pentland 2006] considered z-score as a measurement of speech activity. It provides the probability of a person speaking with respect to the rest of the speakers but requires all users’ speech segments to perform the computation. CoenoFire [Feese et al. 2013] performed a more lightweight speech detection method but considered only the overall speaking time. The initial process of personalized speech detection induces the computational burden, in contrast to speech activity inference, which is mapped to statistical analysis on the speech segments of a specific user.

5.1.5. Auditory in Essence. When Pentland [2007] introduced social signals, he proved the applicability of the domain based on behavioural cues that were extracted from audio data. This induction may undermine the importance of these types of behavioral cues but also may indicate how informative they are. Kendon [1990] refers to conversation as an occasion of social interaction that highlights the correlation of the terms in the context of human behavior. For that reason, a great deal of nonverbal communication may accompany the spoken linguistics. As shown, features related to the interlocutors’ voice can provide significant and informative features. However, some feature extraction processes introduce a considerable amount of computational burden and energy consumption, but there are alternatives that can be utilized. A considerable amount of research has been conducted based on audio data. Thus, applying these techniques on mobile devices and in real-world environments will constitute a stable and robust solution for social behavior inference, with a proportional cost in computation and energy.
5.2. Physical Activity

After the incorporation of an accelerometer sensor on mobile phones, initially by Nokia and later by Apple, which led to the evolution of smartphones, physical activity became a popular behavioral cue [Choudhury et al. 2008]. This provides additional contextual information allowing the reduction of false positives in situations such as stress detection. Furthermore, it may be utilized as an optimization process (e.g., to discard data under the assumption that when the user is running, he or is not socially interacting). Most approaches focused on computing statistical features showing how active the user is but also what kind of activity the user was performing, such as standing, walking, or running. In order to classify the activity the user is performing, researchers mainly train an activity detection model for each category that requires identification. In the next subsections, state-of-the-art techniques are presented that have been developed for measuring how active users are and in which activity state they are in.

5.2.1. Movement Activity. One of the cues a person is conveying when socially interacting is movement activity (e.g., small but noticeable activity when in a standing position, which may reveal social behaviors such as stress). A low-complexity and robust method was presented in Social fMRI [Aharony et al. 2011]. The authors retrieved 3D accelerometer data from mobile phones for 15 seconds every 2 minutes, and then they computed a vector with the frame’s magnitudes. They then calculated the variance for each second of the frame, which was utilized as a ranking mechanism to classify the user’s activity as still, moderate, or high based on a threshold.

In Jigsaw, Feese et al. [2013] extract movement activity by retrieving information from the accelerometer of the mobile phones. First, they calculate the standard deviation of a moving window and then, based on a threshold, they classify the segment as active or nonactive. In addition, they compute the movement intensity through the median magnitude of the linear absolute acceleration. It should be mentioned that these features are focused on firefighters, who work in an intense environment. Berke et al. [2011] utilized the accelerometer and barometer sensors to detect users’ activity and the time spent performing the specific activity given a certain weight. Based on the importance of the activity, they provided an analogous weight to the time a user spent in performing a certain activity. Muaremi et al. [2013] computed activity movement of a user based on accelerometer and GPS data of the mobile device. From 3D accelerometer data, they calculated the magnitude and further mean and variance values. From GPS data, the amount of locations a user visited and the traveled distance were computed.

Discussion. Movement activity detection has been tackled with several methods, mainly by computing statistics such as variance of a window over raw 3D accelerometer data and then applying an empirical threshold. Social fMRI [Aharony et al. 2011] is one of these methods, which can be utilized easily through the Funf [Funf 2011] open-source sensing framework that performs the procedure as a preprocessing step. A similar technique was implemented in Coefire [Feese et al. 2013]; in Berke et al. [2011], where authors included a time-dependent factor; and in Muaremi et al. [2013], where the authors incorporated a location-change magnitude feature. Given the energy consumption of the aforementioned sensors, the most suitable approach for continuous sensing applications is the extraction of accelerometer statistics, which is a lightweight and reliable procedure [Hoseinitabatabaei et al. 2013]. Although, the location incorporation provides an additional measurement and parameter, it is based on GPS: a high-energy consuming sensor that cannot be used for long-term sensing applications or should perform sensing in an efficient manner. However, a less accurate approach may be utilized such as GSM localization techniques to provide coarse-grained location estimation. In essence, detection of activity movement for continuous sensing applications may focus on less power-“hungry” sensors such as the accelerometer and
extract information through low-complexity statistics suitable for the majority of social behavior applications.

5.2.2. User’s Activity State. Activity recognition provides important information for understanding the context in which a social interaction is taking place. Also, it allows researchers to create more accurate and reliable techniques for social behavior understanding based on a specific activity state. There has been a considerable amount of research in order to detect the activity that a user is performing based on sensor data retrieved from mobile phones. In this part, we mention only the most important works analytically, to provide an overall understanding to the reader regarding the process of activity recognition. An extensive survey on activity recognition through body-worn sensors is presented in Bulling et al. [2014].

Yang [2009] presented an activity recognition approach that utilizes orientation-independent features for vertical and horizontal components of accelerometer data. He computed the mean, standard deviation, ZCR, interquartile range, 75% percentiles, spectral entropy, and entropy of both components and their cross-correlation. The features were forwarded to a decision tree [Quinlan 1993], leading to 90% accuracy for sitting, standing, walking, running, driving, and bicycling. Also, he proposes an approach to reduce data overfitting that combines K-Means clustering (a cluster for each activity) followed by an HMM-based Viterbi algorithm to leverage historical data.

A lightweight approach for detecting users’ state is presented in Wang et al. [2009] that is based on the standard deviation of the magnitude of accelerometer data. It does not rely on the device orientation and achieves above 70% accuracy. Initially authors retrieve N (window size) number of measures from the accelerometer and convert them to magnitude time series. Then from the produced signal, they extract entropy, power, value, and amplitude of the highest-magnitude frequency and weighted mean/variance of the top-t highest-magnitude frequencies (weighted by amplitude). For classification, they perform offline supervised training for each user on a C4.5 decision tree model (70% accuracy) by utilizing accelerometer data that are labeled with respect to the activity that is performed. Feature extraction and classification are executed in real time [Srinivasan and Phan 2012].

Jigsaw [Lu et al. 2010] performed activity recognition and divided the process into four stages: (1) calibration, (2) preprocessing, (3) feature extraction, and (4) activity classification. In the first stage, the authors calculate the offset accelerometer parameters for the specific device through a linear least square estimator. The preprocessing stage includes outlier removal and projection of the accelerometer data to the earth’s coordinates. Then, time and frequency domain features are extracted based on mean, variance, mean-crossing rate, and spectral analysis. Finally, a decision tree is created followed by a sliding-window smoothing method achieving 91.64% accuracy for cycling, driving, running, stationary, and walking. Based on this method, they provided an extension that through a crowd-sensing technique creates a personalized model for detecting activity in a population of users [Lane et al. 2011].

Seiter et al. [2013] utilized mobile phones that incorporate accelerometer, barometer, and GPS to understand the level of pain in a patient based on his activity. The authors concluded that based on 40% of pain relief, 10% degradation was detected. It should be mentioned that the study was conducted on only one person. Also, Yan et al. [2012] present an adaptive activity recognition method that leverages mean, variance, entropy, and energy (FFT) of a frame in order to compute the confidence of an activity and then apply the appropriate pair of sampling frequency and feature set. Furthermore, PBN [Keally et al. 2011] describes an approach with multiple sensors deployed on the user. For each sensor, a unique classifier is trained and an overall classifier based on Adaboost [Freund and Schapire 1997] computes the user’s state on the mobile phone.
Discussion. For estimating the user’s activity state, research focused mostly on extracting statistical and spectral features from a window of accelerometer samples and based on these features train a C4.5 decision tree. This constitutes a simple and straightforward approach, with lightweight a feature extraction and classification model, ideal for mobile devices. Regarding the features, some exceptions are the processes of computing FFT and spectral entropy, which add a considerable burden to the device. The Jigsaw [Lu et al. 2010] sensing framework managed a comparable accuracy 91.64% on mobile phones. Both approaches could provide a realistic solution in order to tackle a certain problem. Achieving similar accuracy without the need for extracting the burdensome features could constitute a significant challenge. In PBN [Keally et al. 2011], training a specific classifier for each on-body position and combining those in a hierarchical model provide an accurate outcome but limit the pervasiveness of the system in daily lives. In conclusion, incorporating users’ activity state recognition allows the creation of specific social behavior inference models for different activities. These activity-dependent models led to reducing the error in social behavior inference.

5.2.3. An Outline of Physical Activity. Detection of users’ physical activity on mobile devices is an area that has triggered the interest of scientists mainly from the point of accelerometer integration in COTS smartphones. The works related to physical activity are highly correlated to the detection characteristics of a user’s movement. This includes qualifying that a user is more active than another and classifying users’ current activity state such as standing, walking, or running. Both fields provide important information about the user and his or her behavior but also about the context in which he or she is. To summarize, movement activity detection is a relatively simple and lightweight process that is supplied by several sensing frameworks as a preprocessing step. In contrast, users’ activity state has a great spectrum of inference techniques that adds a notable amount of computational burden and demands model training for preknown states. However, a user’s activity state redounds significantly in detecting the context in which the social interaction is taking place and should be included in the inference process.

5.3. Gesture and Posture

Gesture and posture are two means through which people emit signals during social interactions. A noticeable difference exists among several cultures, but both cues convey important information about the social situation, attitude, relationship of the participants, and so forth. In SSP, gesture and posture inference is performed through video recordings, in which researchers detect certain body parts of the participants. Then, by tracking these points, they train classifiers that infer about various gestures and postures [Vinciarelli et al. 2009]. In this section, we present techniques developed to detect various gestures and users’ posture through mobile phones and in some cases with the incorporation of external sensors.

5.3.1. Gesture. There is lengthy research regarding gesture recognition through several means such as video and body-worn sensors. Literature has focused on utilizing smartphones’ integrated sensors but also incorporating external hardware. Regarding the inference, a popular approach is to train a Markov model for each of the targeting gestures, and based on the confidence that each model produces, the highest is selected. PEYE [Hua et al. 2007] focused on detecting simple motion gestures on mobile phones by utilizing the camera and recording small videos. These are split into sequential images from which they extract small rectangles that are tracked through an adaptive block matching approach in order to understand the device’s movement and, further, users’ hand gestures, with 12.86% minimum matching error. e-Gesture [Park et al. 2011] proposed to train a generic HMM classifier for each gesture, which retrieved data
from a hand-worn sensor and forwarded them to a smartphone. The accelerometer and gyroscope data are segmented in an adaptive manner based on gesture change through their magnitude. Then, they are forwarded to an adaptive (Maximum Likelihood Linear Regression for model update) or multisituation HMM (one HMM for each situation: ride, stand, walk, run) for gesture recognition in four different situations, achieving 84.6% and 94.6% accuracy, respectively.

Junker et al. [2008] extracted the pitch and roll of a user's hand from body-worn sensors. By applying preprocessing mechanisms and an HMM, they were able to identify a user's gestures with 97.7% accuracy. The preprocessing step was based on SWAB [Keogh et al. 2001], which performs segmentation and approximation on time-correlated data. These segments were grouped based on resemblance and the ones with the lowest similarity were selected.

uWave [Liu et al. 2009] introduced Dynamic Time Warping (DTW) that performs adaptive gesture classification through sensed accelerometer data based on only one training sample and achieved 98.6% accuracy on eight gestures. A preprocessing step is completed, which performs quantization on raw accelerometer data to remove noise and reduce the size of the data.

As an improvement of DTW, Wu et al. [2009] presented a frame-based descriptor and multiclass Support Vector Machine (SVM) that was able to detect 12 distinct gestures with 95.21% accuracy. Myo [Myo 2013] is a newly developed wearable armband that is able to perform gesture and motion control. It detects the muscle movement of the user's arm (electromyography) and transmits that information through Bluetooth to another device such as a mobile phone. It should be noted that Myo is a commercial product and its accuracy is not provided. For further information about the analytical works that have been done and proposed techniques in the area of gesture recognition, we refer to an extensive survey [Mitra and Acharya 2007].

Discussion. In essence, PEYE [Hua et al. 2007] performed mainly device movement recognition through video recording. This may be replaced by orientation sensor readings due to lower power consumption of the sensor and process. e-Gesture [Park et al. 2011] proposed an adaptive method that continuously learns based on user labeling but has increased computation due to learning and may not perform well due to data diversity. An adaptive model for each situation (e.g., standing, walking, etc.) may achieve higher accuracy as it will create an activity-dependent classification. The multisituation approach had the highest accuracy but requires training of known situations. Junker et al. [2008] focused mainly on detecting the type of activity the user was performing, while gesture recognition was deficient due to lack of a garbage model: a model that infers if none of the target classes are detected. uWave [Liu et al. 2009] claimed high accuracy (98.6% for eight gestures) through its adaptive approach. However, it should be noted that it is a user-dependent method that must be personalized to each user and there is a requirement for tracking the device orientation in case it is tilted. Myo [Myo 2013] provides a predefined detectable gesture collection, but a developer is also able to add his or her own. Overall, both HMM and DTW methods achieved high accuracies. Because every person is different, a personalized approach will achieve the highest accuracies. However, this constitutes a tradeoff due to the requirement of additional training and user dependence. Also, in e-Gesture [Park et al. 2011], the utilization of a limited amount of training data leads to the computation of nonoptimized thresholds and as a result the adaptive methods may achieve lower accuracy. Overall, approaches that utilize inertial data are more suitable than the video-based approach due to mobility and energy restrictions; however, adaptive sensing and inference may be required.

5.3.2. Posture. A person’s posture is divided into head and body posture. Both produce nonverbal communication during a social interaction through the tilt of a certain body
part. This is a brief description of existing literature for both classes in the following subsections.

5.3.2.1. Head Posture. Being able to detect head posture through a mobile device can provide valuable information about a social interaction, such as where the user is facing and if the head is tilted. For example, during a social interaction when people have a common interest or agree on a certain topic, they tend to tilt their head to the same direction, that is, mirroring [Chartrand and Bargh 1999]. Thus, head posture detection is another significant behavioral cue that could be utilized for social behavior inference.

In SEPTIMU [Hong et al. 2012], an accelerometer and a gyroscope are integrated inside an earphone and the microphone transmission of the headset is utilized to communicate with a mobile phone to infer about the head posture of the user. Smart Pose [Lee et al. 2012, 2013a] employed the orientation sensor and the front camera of the mobile phone to calculate the user’s neck angle with respect to the earth’s coordinates. Initially, the system performs face detection through Android API built-in functionality. It identifies if the user is holding the device in the hand by shake detection (threshold based) on accelerometer data. Finally, based on the orientation sensor and the user's viewing angle with respect to the device, it computes the average neck tilt angle. Another technology that has been developed is Google Glass [Google Glass 2013]: wearable glasses incorporating multiple sensors such as gyroscope, accelerometer, and magnetometer, allowing them to be utilized as nonintrusive technology.

Discussion. Research has focused mainly on obtrusive approaches for head posture detection. SEPTIMU’s [Hong et al. 2012] claim of head tracking provides a simple but obtrusive solution because it requires the user to continuously wear an earphone. Smart Pose [Lee et al. 2013a] constitutes a low-complexity and multimodal approach for head posture detection without any external hardware; it relies on off-the-shelf smartphone-integrated sensors but requires the user to hold the device in the hand and also interact with it. Lately, wearable devices constitute a viable solution for accurate and reliable head posture inference.

5.3.2.2. Body Posture. An equally important class is body posture detection that conveys social signals such as mirroring and intimacy during a social interaction. While people are interacting, they tend to bend toward a person showing a certain level of intimacy, while a body slope opposite to the interlocutor may indicate inconvenience. Thus, detecting these types of signals can provide underlying information about a social interaction.

In imWell [Jovanov et al. 2013], a sensor incorporating an accelerometer is placed under the left arm of the user and it transmits logged data via Bluetooth to a mobile phone. To detect a different posture, its targets identifying transition points. A pre-processing step is applied that computes the standard deviation of a 1-second window of accelerometer samples, to remove minor movements. Then, the angle change with respect to the vertical position is computed, which determines the upper body posture. Having the standing position of the user as a reference, the authors categorize the user’s body posture based on certain thresholds. CONSORTS-S [Sashima et al. 2008] utilizes the average of the accelerometer window of samples from a wireless on-body sensor and based on the device inclination classifies through decision rules about the posture of torsos (standing, facing up or down). Liu et al. [2012] perform body posture recognition by retrieving measurements from a mobile device's orientation sensor and especially pitch to classify if the user is sitting or standing while the smartphone is in the user’s trousers pocket. They allow a margin of error of 20° to around 180° or 0° of pitch to infer that the user is standing. With the same error margin of around 90° or
−90°, they estimate if the user is sitting. Another approach that provides information about users’ torsos’ facing direction with respect to the earth’s coordinates is uDirect [Hoseinitabatabaei et al. 2013], which utilizes inertial sensors of off-the-shelf mobile phones.

**Discussion.** The literature has targeted mainly inertial sensors to estimate body posture. imWell [Jovanov et al. 2013] utilizes a very simple technique to identify different body postures. However, it is considered as an intrusive methodology because it utilizes an external sensor that is tied around the user’s torso. CONSORTS-S [Sashima et al. 2008] performs rule-based decisions on the average of accelerometer sample windows. It constitutes a lightweight process regarding feature extraction and inference but is susceptible to on-body position changes of the device. So, in order to improve the accuracy of the approach, it requires the creation of different rules for each on-body position. Liu et al. [2012] described a threshold-based approach on the orientation sensor’s pitch. It is easily implementable but is applicable only for the trousers pocket and requires the device to be in a vertical position; thus, it does not provide a generic solution. Regarding uDirect [Hoseinitabatabaei et al. 2013], it assumes that the relative orientation between the device and the user’s body is static. In unconstrained environments, the devices are not fully attached to a user’s body and are able to move in a certain range. Also, a preprocessing step is required to identify the on-body position of the device. Overall, the extraction of body posture is mainly based on accelerometer data, a relatively low-energy consumption sensor but one that still requires an adaptive sensing mechanism. Also, the on-body position of the device may be considered as contextual information to target the inference on specific body parts and discard unrelated positions.

5.3.3. **Revealing the Methodology.** Overall, gesture detection on mobile devices in the current literature requires either the user to hold the device in his or her hand or the incorporation of an external sensor. The integrated accelerometer could be considered as the main source of data, through which gesture-specific models can be trained. In posture detection, head tracking solutions are mostly intrusive (video or external sensor); however, body posture detection could be implemented with COTS mobile phones. The increased popularity and close body attachment of wearable devices that connect with smartphones show good potential in real-world situations for both gesture and posture detection. A gap identified in the literature is the lack of on-body position detection of the device before performing body posture inference. The on-body position of the device [Hoseinitabatabaei et al. 2014; Shi et al. 2011] constitutes a necessity in order to accurately compute the posture in real-world applications.

5.4. **Facial Cues**

One of the most expressive parts of the human body that people used to externalize their interest, agreement, disagreement, surprise, and so forth is the face. This emission of social signals is mainly achieved through facial expressions and eye movement. Thus, providing a detection and quantification mechanism of behavioral cues vented from a person’s face is not negligible while relying on mobile devices.

5.4.1. **Facial Expressions.** People communicate verbally during a social interaction and in parallel emit social signals through their faces. Several works in psychology showcase the importance of facial expression in recognizing interlocutors’ emotions such as valence, arousal, disgust, embarrassment, and amusement [Ekman 1993]. In addition,

---

7The most common wearing positions are trousers pockets, belts, hands, chest pockets, handbags, and backpacks [Ichikawa et al. 2005].
due to this high correlation, they claim a high accuracy in detecting these emotions through facial expressions. For a comprehensive literature review of face recognition, the reader is referred to Hjelmas and Low [2001] and Zhao et al. [2003]. State-of-the-art techniques for detecting facial expressions through mobile phones are presented in the following paragraphs.

Detection of facial expression of a user through a mobile phone is presented in Visage [Yang et al. 2013]. The approach is based on information retrieved from the camera and motion sensors of the device, for face and head pose recognition, respectively. Through the data, the authors perform face detection [Viola and Jones 2004] (Adaboost-based object detection), inferring rectangle features allowing face tracking. The knowledge of head and face reference points combined with Active Appearance Models [Matthews and Baker 2004] merged texture and shape of the face and allowed them to detect different facial expressions. Performance of real-time training and recognition of facial features is presented in Choi et al. [2011] based on a novel nonorthogonal local random basis. According to the authors, this method provides a robust but energy-efficient solution for extracting facial characteristics. The features are forwarded to a neural network that performs classification and updates the decision thresholds. They evaluated against six well-known face databases and benchmarked against the Principal Component Analysis (PCA) approach [Turk and Pentland 1991].

Another proposed method in order to detect facial expressions on mobile phones is Cho et al. [2009]. Initially, the authors proposed the utilization of two SVMs, a micro and a macro component. The first layer computes the score of the input image with respect to a pretrained (non)facial classifier and then a second-layer SVM calculates the fiducial points: eye, nose, and mouth. The acquisition of fiducial points leads to the extraction of Local Gabor wavelet features through the Gini selection method. Doukas and Maglogiannis [2010] performed face recognition on mobile phones based on eigenfaces [Turk and Pentland 1991], a well-known and established approach. By retrieving an image from the smartphone’s integrated camera, they detected the user’s eyes. Utilizing the eyes’ position, they were able to mine several facial feature points. By tracking these facial features, a classifier can be leveraged that enables the identification of various facial expressions.

Discussion. Most of the works in facial expression detection were designed, implemented, and evaluated in constrained environments, mainly to showcase the applicability of such cumbersome techniques. One of these works is Visage [Yang et al. 2013], which was able to detect facial expressions on mobile phones through a well-established face recognition method [Viola and Jones 2004]. However, the system realization in real-world environments including energy consumption was not evaluated, implying its limited applicability in realistic situations. Applying the NN approach for classification on mobile phones in Choi et al. [2011] was proven insufficient. It is worth noting that although PCA with SVM outperforms a conventional random basis, it requires training. Cho et al. [2009] tracked users’ eyes, mouth, and nose, which provided some concrete identification points for facial expression detection. During real-time classification, they claim an inference cycle of 5 seconds by the Boosting Naïve Bayesian (BNB) algorithm, with an overall accuracy of 75% for four expressions. Doukas and Maglogiannis [2010] applied eigenfaces [Turk and Pentland 1991], distance projection and computation, and classification on mobile phones. The inference process duration was about 1.2 seconds, which constitutes a reasonable delay for face recognition. However, they have not included any analysis regarding the computational burden and the energy consumption of the approach, which will limit the applicability of the method in real-world situations. Even though facial expression detection constitutes a cumbersome process, the value of this behavioral cue is important.
5.4.2. Eye Tracking. During a social interaction, a person conveys nonverbal communication signals from the eyes. Social signals indicating intimacy, interest, personal relation, and conversation regulation are some of the information emitted through eye contact [Kleinke 1986]. Furthermore, Vertegaal et al. [2001] induce the ability of predicting interlocutors’ attention during a group conversation based on eye movement. Thus, as eye movement is considered an informative behavioral cue in social behavior inference, state-of-the-art techniques for eye tracking are reviewed in the following paragraphs.

One of the first works for detecting eye movement was eyeLook [Dickie et al. 2005], where authors attached an Eye Contact Sensor (ECS) [Vertegaal et al. 2002] surrounded by LEDs on a mobile phone. Through the flashing of LEDs, their reflection was displayed on the user’s pupils, logged by the attached camera, and then transmitted for offline analysis. The LEDs’ reflection was displayed near the user’s pupils which allowed detection and tracking of his eye. By detecting the movement of the participant’s eye, they were able to identify turn taking among users.

For detecting facial movement and especially eye motion, Bulling et al. [2011] developed custom wearable goggles. These were constituted by dry electrodes, a light sensor, and an accelerometer and were connected through a wire to a digital signal processor (DSP) and data storage. Based on the application, they proposed alternative methods, such as electrooculography (EOG) or camera recordings, to detect eye movement. EOG is performed through dry electrodes of the goggles that are attached to participants’ faces. The authors depict six main feature categories from which they compute various statistics and signal characteristics: (1) saccades, (2) fixations, (3) blinks, (4) microsaccades, (5) vestibuloculare reflex, and (6) smooth pursuit movements. It should be noted that the authors utilized only saccade, sequence, blink, and fixation features. By triggering the interest of a participant wearing the goggles, these features are extracted and a model is trained, which is applied later on to identify certain eye movements.

In order to create an eye-controlled mobile phone, Pino and Kavasidis [2012] utilized the front camera of a smartphone for achieving eye tracking. The system takes sequential pictures from the user. Then, it performs a Haar classification that identifies features in a rectangular space through summing the intensity of the pixels. The method detects two similar spaces of the picture and classifies it as the eyes. Having detected the user’s eyes, they utilize the CAMSHIFT [Bradski 1998] algorithm to keep track of their position in upcoming images. To transform the detected eye position to the device’s display coordinates, they compute the centroids of the rectangles and then apply the Starburst algorithm [Li et al. 2005] for tracking.

Discussion. A common practice for extracting facial cues is defined by detecting initially fiducial points [Turk and Pentland 1991; Belhumeur et al. 1997]. Eye positions are some points that may be leveraged for tracking a person’s eye movement. Regarding the aforementioned works, eyeLook [Dickie et al. 2005] is based on an obtrusive mechanism that requires detecting the flash reflection near the eye pupil, which is additionally prone to daylight reflections. Bulling et al. [2011] tried to deviate from the main visual-based approaches by utilizing an EOG attached to muscles surrounding the eye. This is a less computationally consuming approach that provides a raw signal indicating the muscle movement, allowing detection of eye activity. Although the method requires specific glasses with integrated dry electrodes, the reduction of the complexity is noticeable regarding the video-based approach. Furthermore, Pino and Kavasidis [2012] also focused on a video-based eye tracking method. By applying a modified Haar feature extraction and classification, they were able to achieve a speedy inference of eye points in an efficient manner. The classification process was only initiated when a certain movement threshold was overcome. Finally, performing eye tracking requires tackling some key challenges including (1) the high computational
power required for retrieving information from visual data, (2) the difficulty in
detecting the eye pupil in a sunny outdoor environment that is characterized by
brightness fluctuations and saturation [Drewes et al. 2007], and (3) the utilization of
ubiquitous and nonintrusive sensors for retrieving data that will allow the inference.

5.4.3. Facial Cues at a Glance. The face is a very descriptive part of the human body
during social interaction in terms of social signal conveyance. However, the capability of
detecting these behavioral cues on mobile devices constitutes a great challenge. Along
with the computationally complex process of detecting facial cues, other provocations
arise, including the intrusiveness of the system, the applied training data, and the real-
time on-device classification. Detecting facial cues is stemmed by identifying several
fiducial points of a person’s face, including the mouth, nose, and eye. In many cases,
the methods include a preprocessing step of detecting these parts of the face and then
utilizing these points for classifying facial expressions and monitoring eye movement.

Overall, identifying facial cues (i.e., facial expressions and eye tracking) is a burdensome
process, especially for constrained devices such as smartphones. As shown, there are
works that have managed to execute these computationally demanding processes on
mobile phones. However, they do not provide a concise energy consumption and compu-
tation burden analysis that will indicate the applicability in continuous sensing and in-
ference applications. SociableSense [Rachuri et al. 2011] is a system that decides based
on computational requirements to perform the inference on the device or at a cloud
infrastructure. Applying a distributed inference adaptation model such as Sociable-
Sense combined with an adaptive sensing technique based on the context is a viable
solution for performing such burdensome processes on smartphones. Finally, wearable
technologies such as Google Glass [Google Glass 2013] constitute a promising approach
for real-world applications. However, energy consumption is still a great challenge in
continuous sensing systems.

5.5. Environment and Space
Equally important behavioral cues are space and the environment in which a social in-
teraction is taking place. According to psychologists, the interpersonal distance and the
spatial arrangement of the interlocutors provide a large amount of information about
their social relation, their intimacy, and the probability that people are interacting in
multipersonal interactions [Goffman 1963; Hall 1966].

5.5.1. Interpersonal Distance. In psychology, proxemics is an area that has been exploited
for many years, starting from the work of Hall [1966]. In this work, Hall, following the
social behavior among animals, defined some imaginary concentric circles around each
person during a social interaction, which indicates the type of relationship among the
people. Interpersonal distance is a significant element of social interactions, not only
to detect if people are interacting but also to estimate their relationship.

5.5.1.1. Sound. The most promising approach for distance estimation through sound
is BeepBeep [Peng et al. 2007]. It is based on ToA without the requirement of time
clock synchronization among the devices. Each of the devices sends out an audible
beep sound and logs its own sample and the remote sound. The device continues to
record until it receives the remote beep. Then, they exchange the standby time and
compute the interpersonal distance from the number of samples recorded and the time
required to receive them. An extension of BeepBeep is Liu et al. [2013b], in which
the authors develop a transmission scheme and apply an adaptive ToA mechanism
to improve the accuracy of the system. Also, Whistle [Xu et al. 2011] is an approach
akin to BeepBeep but relies on TDoA by recording the sound from multiple devices
and performs the computation at a centralized point. A recent work called RF-Beef
[Uddin and Nadeem 2013] combined the methodology of BeepBeep with an RF interface to apply TDoA by sending initially a beep sound followed by an RF beacon. A ToA-based scheme is introduced in Liu et al. [2013c] that uses a speaker and a mobile phone’s microphone to perform distance estimation. Filonenko et al. [2013] exploited and developed a mechanism for estimating the distance based on TDoA among devices by producing ultrasound through COTS mobile phones.

Discussion. A different modality for distance estimation, the sound, was considered in some approaches in order to be able to apply techniques that are difficult to deploy on mobile phones such as ToA, TDoA, and AoA. BeepBeep was the first work that was able to leverage these types of techniques (ToA) on mobile phones based on sound. By exchanging the time duration, there was no need to perform clock synchronization among the devices. The technique was applied between two devices only. Following this approach, different works used other techniques such as TDoA or combined them with RF signals. As claimed, they are able to achieve a fine-grained distance estimation among the devices. However, the sound-based methods are prone to the relative orientation of the device and user with respect to the interlocutor. The majority of these approaches utilize audible beacons that are not suitable for ubiquitous usage. Filonenko et al. [2013] claimed to have achieved the transmission of ultrasound through COTS mobile phones. For the appliance of TDoA, there is a requirement for speaker array deployment in the environment in order to calculate the time difference between arrival of the beacon at the two speakers. This increases the intrusiveness of the system.

5.5.1.2. RF Interfaces. Interpersonal distance estimation through RF-based technologies (e.g., RF, Bluetooth, WiFi) constitutes a common approach due to its easiness in development and implementation. Researchers have developed various techniques to estimate interpersonal distance among users in a coarse-grained and fine-grained manner.

Coarse-grained. A commonly used method to detect if people are in vicinity is to utilize the Bluetooth interface. This is available in the majority of today’s mobile devices. By performing an enquiry scan process, a smartphone retrieves discoverable nearby devices. This approach takes into consideration every device that is in the range of the Bluetooth radius (\(\sim 10\)m). It is not affiliated with any intelligence to mine more specific information about interpersonal distance; only details such as the identifier and timestamp are logged. One of these works was Serendipity [Eagle and Pentland 2005], in which the author developed a BlueAware framework for mobile phones to log the Bluetooth identifier and current timestamp. Antoniou et al. [2011] deployed Bluetooth dongles inside a building and through Bluetooth discoverable mobile phones they were able to detect if users were in vicinity. CenceMe [Miluzzo et al. 2008] and Friends and Family [Aharony et al. 2011; Staiano et al. 2012; Singh et al. 2013a; Bauer and Lukowicz 2012; Efstratiou et al. 2012] are other examples of works where they utilized simple Bluetooth discovery to infer if users were in vicinity. Unlike previous approaches, PeopleTones [Li et al. 2008] leveraged cell tower readings to estimate if the users are nearby in a larger scale, claiming an error around 322m.

Discussion. As noted before, the most common approach for distance estimation on mobile devices is through RF-based technologies. In coarse-grained distance estimation based on the targeting device, researchers focused on Bluetooth discovery or GSM localization. On one hand, due to the popularity and implementation simplicity of Bluetooth discovery, it constitutes a widely used method when conducting research into social behavior. It consists only of the discovery of nearby devices and logging their BTIDs, including the timestamps. There is no processing or inference required; thus, if the induced error in distance estimation is acceptable for a certain type of
application, this method may be preferred. On the other hand, a large amount of people set their devices on nondiscoverable mode or disable the Bluetooth interface of their smartphones, making the coarse-grained distance estimation nonapplicable. Nevertheless, the range of Bluetooth introduces a large amount of error; for example, two people may be in different rooms, but through this method they are considered close enough to interact. These interpersonal distance estimation techniques rely on the assumption that when devices are in vicinity, then their users are as well. However, in real-world situations, this assumption is not always valid; thus, there is a need to incorporate a mechanism to detect when a user is not carrying the device, such as in Hwang and Wohn [2013].

Fine-grained. Alternative and more advanced techniques have been proposed to achieve a more accurate result in estimating if people are in vicinity. These techniques are mainly based on ToA, TDoA, AoA, and RSSI. Due to ease of implementation on mobile phones, most approaches focus on retrieving the Bluetooth/WiFi RSSI, and then through a PLM, threshold-based classification, or machine-learning technique, they try to estimate the interpersonal distance of the users.

An initial approach to estimate interpersonal distance through Bluetooth/WiFi RSSI is the development of a PLM. The most simple method is Free space PLM, which considers an ideal environment without reflections and obstacles. It requires a reference $RSSI_{ref}$ measurement at a specific distance. Given the RSSI reference, the model estimates the distance between the two devices. An improvement of this model is Office PLM [Wang et al. 2013], which modifies Free space PLM. In particular, it adds the impact of the indoor environment and especially of a normal office while assuming line of sight between the devices. Regarding the environmental parameter for indoor environments, there are predefined values for certain types of rooms that can be utilized. Alternatively, by retrieving RSSI samples at different distances and through an optimization technique, researchers may compute their own parameters. Based on these generic PLMs, several variations have been proposed that add more parameters in order to consider other factors. One of these variations is BlueEye [Ghose et al. 2013], which strengthens the office PLM by incorporating two environmental constants and the relative orientation of the two devices; one of the factors that affects the RSSI is the directionality of device antennas. The output of the improved PLM was forwarded to k-means clustering to estimate users’ interpersonal distance.

Stankovic [Li et al. 2013] applies a PLM with computed parameters for indoor and outdoor environments to detect when people are in vicinity; the interpersonal distance boundary utilized is 3m. Regarding the WiFi interface, Matic et al. [2012] created an artificial dataset through an indoor PLM by leveraging WiFi RSSI measurements at 1m distance. Then, he trained a Naive Bayes with a KDE classifier to detect if people were at a distance to socially interact. Features utilized to train the classifier were average and maximum values of a 10-sample window. Finally, Comm2Sense [Carreras et al. 2012] followed the same process for training a classifier on a 20-sample window that determines in which interaction zone people socialize.

Discussion. Researchers managed to achieve an improved accuracy in distance estimation. In order to achieve this, techniques such as RSSI, ToA, TDoA, and AoA were utilized. For the implementation of some of these methods on smartphones, there are particular requirements such as firmware modification, multiple antennas, and so forth. Thus, research has focused mainly on leveraging RSSI provided from the core API of the majority of COTS mobile devices. Based on RSSI, various PLMs have been proposed for environments such as free space, indoor, and office, which require certain parameters for the specific environment. Even given the environmental parameters, RSSI is prone to antenna type and orientation, human body absorption,
reflections, and obstacles. Banerjee et al. [2010] and Matic et al. [2012] strive to tackle this through machine-learning techniques. They incorporated uncertainty in distance measurements and utilized a 20-sample window on which certain statistics were computed. However, they performed only small-scale experiments, and viability, reliability, and robustness of such a solution in the real-world environment are not proven. Additionally, the number of samples (window) required should be taken into consideration. As the number of samples increases, depending on the sampling frequency of the RF interface, the waiting time for an inference may increase. Also, when using a large window of samples (e.g., 20 samples [Carreras et al. 2012]), the data may be outdated, leading to erroneous inference results.

5.5.2. Spatial Arrangement. Kendon [1990] introduced F-Formation, referring to the spatial formation created by the participants during a social interaction. In more detail, an F-Formation can include various configurations such as face to face, side by side, rectangular, circular, semicircular, and L-Formation. So, depending on the formation that participants frame, different information about their social relationship is conveyed. This signifies the importance of a user’s spatial arrangement. Researchers in Virtual Compass [Banerjee et al. 2010], by considering the interpersonal distance among users in vicinity, created a virtual map through computing the Euclidean distance of the users. Matic et al. [2012] utilized off-the-shelf mobile phones to detect social interactions. Each participant carried the device on a static body position. While knowing the position of the mobile phone relative to a user’s body, they used the orientation of the mobile phone in order to detect forward direction of torsos, and hence to estimate users’ spatial formations.

Discussion. Virtual Compass [Banerjee et al. 2010] calculated users’ relative spatial arrangement. Due to the lack of users’ facing directions and absolute locations, the authors were not able to estimate the absolute spatial arrangement of the users. As the approach incorporated both WiFi and Bluetooth RSSI to perform the computations, unless an energy-efficient mechanism is added, this work is not suitable for continuous sensing applications. Also, RSSI is highly dependent on the environment and prone to human body absorption. On the contrary, Matic et al. [2012] used the orientation sensor to keep track of the user’s facing direction. However, the orientation sensor is based on a fusion mechanism of accelerometer and magnetometer that is affected by accelerometer bias and magnetic disturbance. A fusion mechanism that incorporates a gyroscope with a drift compensation approach could prove to be a more reliable solution. Researchers in this work increased the intrusiveness of the system by limiting the smartphone’s wearing position to the user’s belt. A less restrictive approach regarding the user’s wearing position would improve the pervasiveness of the system.

5.5.3. A Disclosure of Environmental and Spatial Cues. The environment and space in which a social interaction is taking place conveys information. A brief comparison of state-of-the-art techniques was presented to understand and provide quantification mechanisms to allow the extraction of these types of information.

Interpersonal distance estimation is an explored field with several proposed approaches. The classification of these works was based on the modality utilized to perform distance estimation. Sound-based distance estimation is the most recent approach where scientists have shown interesting results. BeepBeep [Peng et al. 2007] was able to tackle the device synchronization problem required in ToA-based methods. Audible beacons constitute an issue that could be tackled through ultrasound beacons; however, they are still in an immature phase regarding mobile phones. In RF-based approaches,
there is high dependence between the accuracy and system complexity required. Techniques such as ToA, TDoA, and AoA are mainly contingent on external stationary or mobile hardware, which introduces a certain level of intrusiveness and also mobility issues. RSSI is a popular solution for estimating distance but is highly dependent on the environment and is characterized by large fluctuations. Overall, these methods are prone to the environment and to human body absorption, which both introduce a considerable amount of error. Preliminary results have shown that ultrasound methods could achieve accurate distance estimation. However, to our knowledge, there is no evaluation in unconstrained real-world environments.

Regarding the spatial arrangement of the users, its importance has been indicated in psychology [Kendon 1990]; however, there is not a considerable amount of work. Researchers focused on detecting the relative spatial arrangement of the users. Furthermore, relative spatial arrangement induces error, as the absolute position is not known and through various parameters researchers focus on reducing the location uncertainty. To our knowledge, there is no analytical work in order to quantify the error induced by this approach. Absolute positioning systems may reduce the error introduced by estimating relatively the spatial arrangement of the users. This could be achieved by inertial tracking systems that are built upon these types of information bearing in mind the requirement for energy efficiency due to continuous sensing.

5.6. Device Usage

The term “behavioral cues” mainly refers to nonverbal signals that are conveyed from a person during a social interaction. This constraint refers not only to physical presence but also to a social interaction in different physical places. For example, during an SMS text conversation, people emit social signals such as response time, call frequency, punctuation, emoticons, and so forth. These are all a small part of features that could be extracted from the usage of a mobile device.

_SenseMs_ [Amin et al. 2005] was one of the first works that argued about nonverbal signals in SMS messaging. Falaki et al. [2010] logged users’ interaction with the device in order to understand the effect on the network and the energy consumption. The data utilized in this work could be forwarded to a human behavior understanding mechanism to extract contextual information. Bauer and Lukowicz [2012] monitored calling and SMS text behavior on the mobile phone of a person and categorized it to different social groups. Li et al. [2013] utilized GPS and calendar to understand the context of a social interaction while logging call records to a list the interlocutors. Altshuler et al. [2013] introduced six categories of features based on users’ patterns that could be retrieved from a mobile phone: (1) Internet usage, (2) calls, (3) SMS messages, (4) phone applications, (5) alarm clock, and (6) Location. _BeWell_ [Lane et al. 2012] also monitored smartphone usage such as device charging, screen lock, power off, and so forth. Apart from the previous works describing the features that could be extracted from mobile phones, Oliver [2010] created a dataset from 17,300 Blackberry devices in which he logged data representative of the user’s interaction with the device. These datasets could prove to be a useful mean for predicting users’ context.

**Discussion.** The most important advantage of these types of signals is that they are collected from virtual sensors. This type of information is stored locally on the device while a person uses it and can be retrieved at the user’s discretion. Researchers can collect these types of data through the device’s API or a sensing framework. Then, they can extract behavioral information with negligible energy consumption due to lack of harnessing any of the burdensome physical sensors. These types of cues can be employed for long-term behavioral analysis of a user by inferring social characteristic patterns, but also to acquire contextual information.
5.7. Physiological

Extracting physiological characteristics of people during social interactions provides precious intelligence of the natural state of the body. During a social interaction, based on the user's mental state, feelings, stress, and so forth, the physiological body states are changing such as heart rate, skin temperature, and humidity. For example, people are interacting and due to the conversation context they feel stressed, which increases their heart rate and skin temperature. To detect these types of signals, researchers have focused on galvanic skin response (GSR), respiratory inductance plethysmography (RIP), electrocardiography (ECG), and electroencephalography (EEG) sensors.

*AutoSense* [Ertin et al. 2011] is a system composed of physiological sensors such as GSR, RIP, ECG, a mobile phone, and a software component called FieldStream. Through external sensors (RIP and ECG), FieldStream performs a windowing pre-processing, producing information such as window of R-peak locations, followed by feature extraction computing mean, variance, heart rate, and respiration rate. *NeuroPhone* [Campbell et al. 2010] was the first work that incorporated mobile phones with a wireless EEG headset in order to perform actions on the mobile device emitted directly from a person's brain. The headset transmits data to the mobile phone, on which an initial averaging is performed followed by the application of a bandpass filter for noise reduction. Then, they utilize weighted classifiers, multivariate equal-prior Bayesian, and decision stump classifiers. This approach could be applied in order to detect other brain signals, which will lead to other social signal detection. *imWell* [Jovanov et al. 2013] connected a smartphone with a physiological sensor called Zephyr BioHarness 3 [Zephyr 2012] through a Bluetooth interface. The mobile phone was monitoring and storing information about a user's heart activity and later uploaded the data to an mHealth back-end server for offline processing. *SEPTIMU* [Hong et al. 2012] utilized an earphone in which a microphone was incorporated in order to detect a user’s heart rate.

**Discussion.** The literature has mainly focused on detecting a person’s heart rate and skin temperature. This is performed through off-the-shelf sensors transmitting through wire(less) communication to mobile phones, which conduct the inference. Off-the-shelf sensors have incorporated mechanisms of noise reduction, and thus provide accurate estimations and usually do not need any preprocessing step. However, current approaches introduce a certain level of intrusiveness, which should be considered during the design of Mobile SSP applications.

6. FROM SOCIAL SIGNALS TO SOCIAL BEHAVIOR INFERENCE

Extraction of behavioral cues constitutes an abstraction layer, in which some preliminary knowledge is retrieved from raw sensor data. Combining these different types of information leads to the process of mining social signals. These signals convey significant information that characterizes a person’s feelings, mental state, interest, and boredom during a social interaction. As the duration of the social signals is limited, a long-term analysis of the information they provide will infer a person’s social behavior. In this section, we will outline different social behaviors that can be extracted from long-term analysis of certain social signals with respect to the behavioral cues analyzed in the previous section (see Figure 5).

6.1. Stress

A social behavior that has attracted a noticeable amount of interest among researchers is stress. Stress detection is mainly based on vocal, physical, and physiological activity
cues that are forwarded to a machine-learning technique responsible for providing an estimation. As claimed by researchers, state-of-the-art techniques are able to achieve an acceptable accuracy of over 80% in most cases.

**AMMON** [Chang et al. 2011] extracted prosodic features and utterances that were fed in a linear SVM and performed stress classification with 84.4% accuracy and 93.6% for stress increase-decrease. **StressSense** [Lu et al. 2012] exploited three different approaches to train two GMMs for stressed and neutral voice. The authors developed a universal model for all participants (71.3% indoor accuracy), an adaptive model that starts from the universal model and through maximum a posteriori fits to a specific user (81.3% supervised, 77.8% unsupervised indoor accuracy), and finally a personalized model trained especially for each participant (82.9% indoor accuracy). **AutoSense** [Ertin et al. 2011] requires physiological measures such as cardiovascular and respiratory data to infer about users’ stress levels with 90% accuracy.

Gaggioli et al. [2012] combined activity, posture, and physiological features through neural networks and a fuzzy logic algorithm in order to detect if a person is stressed. Sun et al. [2012] performed stress classification in three different activities. They utilized physiological (ECG and GSR) and activity (e.g., sitting, walking, standing) features to determine if a person is stressed by applying J48 decision tree [Quinlan 1993], Bayesian Networks [Friedman et al. 1997], and an SVM [Cortes and Vapnik 1995], achieving corresponding accuracies of 92.4%, 85%, and 84%. Muaremi et al. [2013] integrated physical activity, auditory, phone usage, and heart rate variability features and achieved 61% accuracy for stress detection through multinomial logistic regression.

**Discussion.** As shown, existing literature has focused on inferring stress through auditory, activity, and physiological cues. **AMMON** [Chang et al. 2011] was able to manage 84.4% accuracy through prosody including glottal features and utterances given the tradeoff of computational burden introduced by eigenvalues solving and other glottal features. In **StressSense** [Lu et al. 2012], as expected, the personalized classifier achieved the highest accuracy. But for each user there is a need to train a separate model, followed by a supervised adaptation model, an unsupervised adaptation model, and lastly a generic classifier managing the worst accuracy. It is worth noting that external equipment was required in order to be able to perform speaker segmentation, that is, an indoor array of microphones and outdoors a second smartphone.
AutoSense. Ertin et al. [2011] and Gaggioli et al. [2012] require additional physiological equipment. This introduces a certain amount of intrusiveness but includes supplementary features such as heart rate achieving multimodal inference. As opposed to AutoSense, which utilizes a J48 classifier, which is prone to overfitting, Gaggioli et al. [2012] apply fuzzy-logic-based rules that insert softer boundaries in the classification process. Sun et al. [2012] with similar modalities achieved a relatively robust approach, without auditory cues, as for different types of classifiers there is a small variation in the overall claimed accuracy. Muaremi et al. [2013] utilized lightweight and easy-to-extract features but achieved the lowest accuracy for stress detection in the literature we reviewed.

In essence, the approaches for stress detection are concentrated either on auditory cues or on a combination of physiological, activity, and auditory cues. The literature indicates that the most significant cues are auditory and physiological for detecting stress. In detail, researchers were able to detect stress accurately (over 80%) by utilizing only auditory data and extracting the aforementioned cues, in contrast to physiological cues that were combined with additional modalities. Another important point that should be taken into consideration is the identification of the activity that the user is performing before executing the stress classification. Depending on the activity, the approach may be prone to false positives when carrying out intense activities. In conclusion, stress detection is a promising area and with the incorporation of the field of psychology will become mature, multimodal, and coherent.

6.2. Emotion

After analyzing existing techniques for stress detection in Mobile SSP, in this subsection, we focus on emotion detection. To detect emotion in a preliminary stage, researchers perform some simplification by focusing on the identification of major emotions such as happiness, anger, neutral, sadness, and so forth or just classifying if the user has positive or negative emotions. For emotion detection, scientists utilized audio datasets targeting different emotions and used them as training sets for machine-learning techniques. Next we will present state-of-the-art techniques researchers utilized and provide a brief discussion about them.

At first, AMMON [Chang et al. 2011] extracted prosodic and spectral features from the Belfast Naturalistic Database [Douglas-Cowie et al. 2003] and trained an SVM [Cortes and Vapnik 1995] classifier with 75% accuracy for emotion recognition (i.e., positive or negative). An important work is EmotionSense [Rachuri et al. 2010], which used the Speech and Transcripts library [Liberman et al. 2002] to train an emotion recognition model and succeeded in 71% accuracy for five emotions based on prosodic features. Visage [Yang et al. 2013] detected users’ emotion on mobile phones through facial expression detection [Belhumeur et al. 1997]. To evaluate their approach, the authors applied it on the JAFFE dataset [Lyons et al. 1998], achieving the following accuracies: (1) anger, 82.16%; (2) disgust, 79.68%; (3) fear, 83.57%; (4) happiness, 90.30%; (5) neutral, 89.93%; (6) sadness, 73.24%; and (8) surprise, 87.52%. Cho et al. [2009] apply facial expression classification to detect a user’s emotion and discriminate among four different emotions: (1) neutral, (2) joy, (3) sadness, and (4) surprise.

Discussion. As mentioned earlier, AMMON focused only on extracting information regarding the users having a positive or negative emotion, which induces some generalization. By performing classification with several feature sets, they achieved acceptable accuracy given the tradeoff of computational load when including glottal timings in the feature set. Formant tracking including the Newton-Raphson method is a high-workload process, while in case the eigensolver fails, additional burden is created by the construction of Toepliz matrices. FFT is another technique that is computationally

ACM Computing Surveys, Vol. 48, No. 4, Article 57, Publication date: March 2016.
expensive and should be considered before being applied on a mobile device intended for continuous inference.

*EmotionSense* includes components for adapting the sensing process based on the context. It showcases the effects on computation, communication cost, and energy for performing the computations on the device or on a back-end server. The authors trained the emotion detection model on a state-of-the-art library. However, there is a need to evaluate this model not only based on the trained library but also in a real-world environment to understand the robustness of the model. They performed speaker recognition on samples retrieved from 10 users. But there is no indication on what type of environment the data were collected from, that is, indoor, outdoor, with(out) ambient noise, and so forth. Furthermore, adding Brownian noise is not sufficient to prove that the detection model is able to tolerate noise introduced by real-world environments. Similarly, the emotion recognition model was only evaluated on data from the training library. In essence, providing an evaluation of each of the components (speaker, emotion recognition) individually and as a holistic approach on real-world data would indicate the robustness of the system in daily life monitoring. This necessitates the conduction of a larger-scale experiment for further analysis.

*Visage* utilizes a well-established, robust, and accurate method for face recognition combined with the device’s orientation. However, this approach requires the user to hold the device in a position so the mobile phone’s camera is targeting the user’s face. The face recognition approach through Fisherfaces [Belhumeur et al. 1997] provides tolerance in variations of lighting and expressions in comparison to other techniques such as Eigenfaces [Turk and Pentland 1991]. Also, it should be noted that the system operates in a supervised manner. Thus, it requires the user to provide predefined facial expressions to construct a personalized model that classifies the seven distinct emotions.

Cho et al. [2009] were able to achieve a reasonable emotion recognition accuracy (70–80%) for four emotions. They utilized a boosted Naive Bayes for classification that introduces a certain computation load in the training process due to the creation of domain-specific classifiers. Likewise, this approach is prone to the creation of domain-specific classifiers for possible outliers, inducing overfitting. The system requires preloaded images in the device and does not support real-time recognition of users’ emotions through facial expressions.

Based on these techniques, certain parameters should be considered. The highest accuracy was achieved through facial expression recognition in *Visage*. However, it induces intrusiveness due to the requirement that the device’s camera should target the user’s face. Also, the computational burden induced by face recognition and facial point tracking must be considered. *EmotionSense* managed an acceptable accuracy in an energy-efficient manner, without requiring a specific on-body position of the device or any external hardware. *AMMON* provided only a preliminary classification result regarding the user’s emotion but based on the application could be utilized. Regarding Cho et al. [2009], the restricted inference context of the application indicates it as a less qualified system with respect to the others for continuous sensing and inference.

### 6.3. Mood

In contrast to emotion, mood constitutes a generic emotional state difficult to describe and infer due to its multidimensionality. For that reason, researchers tried to approximate this emotional state through detecting several social signals based on their extraction complexity and significance with respect to mood. In order to detect the emitted social signals, researchers employed physiological sensors connected to mobile phones, on which they performed the inference. The most common social signals to infer mood in the literature were arousal and valence, while activeness and pleasure...
were also leveraged. Next, we will outline the techniques developed in the literature and finally provide a brief comparison.

One of the first pieces of research in which the authors interpreted users’ mood, through a sensor that was measuring pressure and arbitrary movement (gestures) the user was applying on it, was eMoto [Sundström et al. 2005]. They decoded valence, effort, pleasure, and arousal. Another work of mood inference on mobile phones was Gluhak et al. [2007]. The authors extracted physiological features from the user and through a certain threshold were able to detect the level of arousal of the user, that is, activated and relaxed. MoodScope [LiKamWa et al. 2013] is a mobile application that takes advantage of a user’s phone usage patterns. Through a 2-month training period, they were able to estimate mood (i.e., activeness and pleasure) with 93% accuracy through multilinear regression. Authors in eyeLook [Dickie et al. 2005] leveraged the ECS eye tracking tool to extract social signals such as attention though fixations and arousal through eye contact.

Discussion. The majority of works have concentrated on detecting mood, especially valence and arousal, based on physiological features. Sensors measuring these types of features provide valuable information about a user’s physiological state but require intrusive equipment that reduces the ubiquitous character of Mobile SSP. eMoto obliges the user to hold a stylus and does not consider the cultural background of the user (e.g., people around the Mediterranean Sea tend to utilize many gestures during their conversation, in comparison to people in Scandinavia, who seldom perform gestures while conversing). Gluhak et al. [2007] considered only the level of a user’s arousal, while the threshold-based classification approach is prone to misclassification when applied to people of different cultures.

As a continuation, MoodScope performed long-term analysis and included also phone usage data. The model’s training required a considerable amount of time. Initial models had a poor performance (60–70%) and only a personalized model was able to achieve high accuracy (93%) in mood inference. Another disadvantage was that the mood detection model needed to be stored at a cloud infrastructure, requiring continuous Internet communication and adding a noticeable burden on battery consumption. It is worth pointing out that the inference is based only on phone usage data, inducing minimal sensing energy consumption. eyeLook detected eye pupils and when they are dilated, which is considered as an indication of arousal. However, the eye-tracking mechanism is quite intrusive and is prone to false positives (eye detection) in an outdoor environment when it is sunny.

The emotional state of mood is not fully described by valence and arousal, indicating that there is a requirement for incorporating other social signals to provide a more holistic approach. A missing parameter is the collaboration with psychologists, who will indicate the grammar of several modalities. This will provide an understanding of the appropriate combination of social signals for inferring mood accurately given its multidimensionality.

6.4. Personality Traits

Following the emotional state of mood, a more static approach in terms of time is the characterization of a person’s personality traits [Goldberg 1993]. These are mainly parts of a person’s character where a long-term analysis is required to identify them. Due to a broad spectrum of personality traits, the majority of researchers have focused their works on the so-called Big Five in psychology: (1) extraversion, (2) emotional stability, (3) conscientiousness, (4) agreeableness, and (5) openness [Barrich and Mount 1991]. Although some works substituted emotional stability with neuroticism, the overall concept of Big Five was the same.
Thus, Chittaranjan et al. [2011] monitored proximity among people and smartphone usage. For each trait, a distinct set of features was fed into an SVM [Cortes and Vapnik 1995] and C4.5 [Quinlan 1993] classifiers to designate the Big Five with accuracies in the scale of 69% to 75.9%. Vinciarelli et al. [2012b] initially examined the correlation of auditory cues with personality traits and then showed that laughter and back-channel influence significantly increased the perception of social attractiveness. Bogomolov et al. [2013] utilized data retrieved from mobile phone usage (calls, SMSs) and proximity (Bluetooth) to classify the happiness of the user with an accuracy of 80.81% through a Random Forest classifier.

**Discussion.** Given the previously analyzed tradeoff of proximity-based detection of social interactions (see Section 4), Chittaranjan et al. [2011] and Bogomolov et al. [2013] utilized simple Bluetooth discovery in order to measure the social interactions in which a user participated. Although this method is easily implementable, it introduces a noticeable amount of false positives that should be taken into account. A supplementary social interaction feature is the number of remote communications that existed among the users, that is, call and SMS logs. These features assume that the owner is the only user of the device and therefore there is a need to immunize it. Overall, the achieved accuracy in both works is acceptable. However, there is a lack of incorporating several informative cues such as auditory and activity-based cues, which in our opinion would provide a significant amount of information about the personality traits of the user.

In contrast, Vinciarelli et al. [2012b] concentrate on auditory cues and show the correlation between them and the Big Five. However, they do not make any attempt to classify personality traits given these specific behavioral cues. The audio data are retrieved from recorded calls and do not include any data from real face-to-face situations. Furthermore, they extracted a large amount of features; some of them are computationally demanding, rising issues regarding the applicability of such continuous feature extraction on mobile phones. Social attractiveness inference is based on laughter and back channels, which were proven reliable cues according to the authors. Other cues could provide additional information such as physical appearance, eye contact, mimicry in speech and movement, and so forth.

In conclusion, inferring personality traits requires mining several social signals. The Big Five is a first step for identifying the most important social signals related to one's personality. Nevertheless, the literature includes works for distinct cues; thus, an initiation of incorporating these different cues will gather a large amount of information and may provide a more holistic characterization of a user's personality.

### 6.5. Dominance

After the analysis of inferring a user's personality traits, another characteristic of social behavior is dominance. During a social interaction, a dominant person has higher social status in contrast to other submissive people. Dominance detection is a popular topic in SSP, which triggered the research on mobile devices as well. In Mobile SSP, researchers mainly inferred dominance through auditory features by applying various distinct sets. In the following paragraphs, the existing literature of dominance inference on mobile phones will be described and analyzed.

We present Aran and Gatica-Perez [2011], which is not based on mobile phones, but the methodology according to the authors is applicable to smartphones. In detail, they propose several approaches, including simple rule-based inference. To introduce multimodality, they perform feature fusion based on the rank or a score and then utilize a rule-based classification. The features are extracted from audio (prosodic and turn-taking) and visual data. MeetingMediator [Kim et al. 2008] also detects dominant persons by computing turn taking (speaking time, average speech segment length),
prosodic features (variation in speech energy), and physical activity. A recent approach was introduced on SocioPhone [Lee et al. 2013b], in which they extracted prosodic and turn-taking features and fed them in a supervised SVM [Cortes and Vapnik 1995].

Discussion. Regarding Aran and Gatica-Perez [2011], the approach utilized in this work constitutes a lightweight and simple method; however it utilizes only one feature. For that reason, the authors decided to perform multimodal fusion. In detail, they applied fusion techniques based on rank or score to generate a unique feature that incorporates a series of multimodal features. In MeetingMediator, the only inference they perform is to compute the correlation of each person with respect to dominance, without developing a dominance detection model. Their conclusion about important features for dominance detection showcases high significance in the speaking time and speech energy variation. Thus, it indicates that a possible dominance detection model should include the aforementioned features. Finally, SocioPhone created an SVM-based dominance detection model but did not perform any evaluation to quantitatively understand the accuracy of the model. It should be noted that they were able to achieve high accuracy in the extraction of prosodic and turn-taking features in different environments and on-body positions of the device. Thus, a real-world evaluation would provide significant information about the applicability of such a model.

6.6. Other Social Behaviors
In previous subsections, we analyzed various social signals that contribute to the inference of some major social behavior characteristics such as stress, emotion, mood, personality traits, and dominance. Based on the literature, these are the main social characteristics that have driven researchers’ interest. However, in parallel with social behavior inference in these works, other social signals were mined that could trigger the interest of researchers to focus on other social behaviors or even invigorate existing inferences.

Other social behaviors were predicted in Singh et al. [2013b], such as diversity (69% accuracy), loyalty (69% accuracy), and overspending (71% accuracy) through phone usage information based on calls, SMSs, and calendar. Berke et al. [2011] calculated the sociability of a person based on the time speaking during his or her participation in a conversation. In SocioPhone, other than training a dominance detection model, they focused on estimating characteristics such as interactivity through the number of turn takings per minute, sparseness based on the number of silences with duration at least 3 seconds, and skewness based on standard deviation of turn takings.

Referring to a previous analysis [Singh et al. 2013b], they were able to achieve a medium accuracy based on survey and receipt/credit data combined with proximity and phone usage data. The features were calculated on data collected over 1 year. Each social behavior considered multiple modalities except overspending, which utilized only proximity data fused with survey data. This method includes survey and receipt/credit data, which induces human error. However, the integration of NFC technology allowing payments through mobile devices combined with incorporation of a connection of the system with a user’s bank account will eliminate the human factor and create an opportunistic sensing system with higher accuracy.

Berke et al. [2011] estimated sociability through auditory data in comparison to SociableSense [Rachuri et al. 2011] that combined speaking time with proximity data. The utilization of multiple modalities allows the inference of a larger amount of information, such as colocation. In the case of an adaptive sensing system, proximity data can be utilized as a mean that triggers the conversation detection module. Thus, there is no need for continuous speech detection while avoiding missing events. In addition, other modalities could be incorporated for sociability inference such as calls, SMSs, and
instant messaging services. Finally, SocioPhone extracted with high accuracy prosodic and turn-taking features, but similar to dominance inference, they did not evaluate their models for interactivity, sparseness, and skewness in real-world situations in order to understand their applicability.

7. APPLICATIONS

In previous sections, state-of-the-art techniques were analyzed that may be used to infer social behavior on mobile phones. Currently, we will showcase the leading application areas in which Mobile SSP can contribute or has already been utilized, indicating the importance and applicability of the field. Among a wide variety of applications where Mobile SSP can be leveraged, we have identified the main areas of healthcare, organizational engineering, and marketing.

7.1. Healthcare

Healthcare constitutes one of the most significant applications of Mobile SSP. A mobile device, through the large variety of internal and external sensors, allows constant monitoring of a patient in an unobtrusive way by simultaneously minimizing the error introduced by human observer. They are able to detect minor and unnoticeable changes or anomalies in behavior, which may lead to diagnosing a disease even in the preliminary stages. Social-behavior-aware mobile devices are capable of benefiting from the diagnosis and prevention of both physical and mental diseases [Pantelopoulos and Bourbakis 2010].

A notable amount of applications focus on the physical illness aspect of healthcare, diagnosis, prevention, and even prediction of various physical diseases. In detail, through continuous monitoring, a minor behavioral variation that may not be noticeable to a human observer or even the patient him- or herself may be identified by anomaly detection in a patient’s social behavior. As an example, Seiter et al. [2013] observed the pain relief of a patient resulting from surgery by detecting behavioral cues such as activity and posture. An application focused more on prevention was presented in Aharony et al. [2011], where a user’s activity was inferred and combined with a reward system to engage users in a healthier way of life. Also, there are situations where the patient requires long-term monitoring of physiological cues such as heart rate, skin temperature, and so forth. These may provide more detailed information about the overall health of the user and predict common diseases such as obesity and high blood pressure and others including multiple sclerosis and Parkinson disease.

Apart from diagnosing physical diseases, Mobile SSP has applications in monitoring mental health as well. This area is described by changes and abnormalities outlined in patients’ behavior that can be identified through continuous monitoring of user behavior. A common application of Mobile SSP is the quantification of users’ stress levels in pursuance of limiting the effects of long-term high-stress levels. This constitutes an application that requires short-range monitoring. However, there are other mental diseases that require long-term monitoring. An example of this is the detection of evidence referring to the possibility of a person being depressed by collecting information such as mobility patterns, sociability, and so forth.

Physical and mental diseases require a continuous, pervasive, and ubiquitous monitoring tool that will provide significant information about anomalies or routines in a user’s social behavior. This will provide unbiased information to medical experts, enabling them to perform an initial diagnosis that will be verified by them.

7.2. Organizational Engineering

Another important application field is an automatic manner in which to quantify and analyze several aspects of organizational engineering, that is, employees’ sociability,
stress, and job satisfaction [Das et al. 2010], including information flow, workload efficiency, and so forth. These are all significant parameters that contribute to a healthier environment with respect to the employee and the organization itself but also increase the efficiency and productivity of the organization.

Mobile SSP will fabricate a new era in understanding, modeling, and predicting the behavior of organizations while introducing the importance of the social aspect. The social behavior of an organization’s employees is an important parameter that is neglected today. So providing a quantification method for employees’ sociability, stress levels, and so forth will indicate the job satisfaction employees are feeling, and accordingly the appropriate adjustments can be performed. As has been shown in preliminary research [Olguín et al. 2009], it is achievable to comprehend the overall workflow at an organization by spotting lack of communication among different departments. This may lead to the identification of any existing or future eruption. Organizations are keen on being knowledgeable about the relationship among people to reduce customer churn (e.g., churn prediction [Dasgupta et al. 2008; Zhang et al. 2012]), to minimize any gap in the functional process of the corporation [Olguín et al. 2009], or to procure a suitable working team [Pentland 2012].

Organizations are dependent on their employees. This indicates the importance of being aware of their healthy social behavior [Bateman and Organ 1983] to cope with early-identified issues such as lack of intercommunication among various teams. Additionally, Mobile SSP may also identify possible unsatisfied customers and further assist in a correctly structured organization. In conclusion, Mobile SSP has the potential to provide various enablers in the field of organizational engineering.

7.3. Marketing

Finally, an area in which several applications of Mobile SSP we foresee will emerge is marketing. Social sciences have become essential in marketing, due to the comprehension of human behavior required to fulfill appropriate needs [Gardner 1985]. The knowledge of the user’s general but also present social behavior constitutes a new parameter in the area of marketing [Chen et al. 2009].

One of the benefits of Mobile SSP is the ability to provide a personalization aspect in today’s generalized marketing campaigns. This allows the identification of certain perspectives of a user’s behavior. Following, it will enable marketers to target their campaigns to a specific audience that is keen on or open to the promotional target [Adomavicius and Tuzhilin 2011]. As an extension, modeling users’ social behavior through a mobile device may guide marketing to a new era, in which the environment will adjust automatically based on a user’s predicted preference and mood [Strohbach and Martin 2011]. Another application that would provide benefits is the identification of potential customers [Gorgoglione et al. 2011]. An example proposed by Pentland [2007] was through leveraging only characteristics of voice, and they were able to predict negotiation outcomes. This achievement would constitute an enabler for telephone-based marketers. In particular, they will recognize quickly customers willing to accept an offer, reducing the time and effort spent for customers unwilling to be convinced.

Overall, as the field of marketing is largely correlated with the area of psychology, there is a large amount of applications that may benefit from Mobile SSP in order to improve and facilitate customer understanding and personalized marketing.

8. DISCUSSION

Mobile SSP is an important domain that has started to gain a great deal of interest due to its wide applicability. There are several fields, not only psychology, that will potentially benefit from the growth of this area. As described, research has not concluded the terms and the taxonomy of SSP. Thus, researchers need to agree and finalize
the terminology of the field so a concrete area is created. This will directly affect the
development of Mobile SSP, while also enhancing the modeling of social signals. Hav-
ing modeled social signals will provide a clearer understanding and classification of
which behavioral cues can lead to certain social signals. By analyzing these signals, an
explicit guideline will help researchers in mining social behavior in the long term.

As shown in Section 3, there are numerous works that have been released in order
to provide the appropriate abstraction for retrieving information from mobile device
sensors. Some frameworks have reached a certain maturity. This enables the utilization
of these tools in the design and development of mobile social behavior applications
without the need for handling low-level procedures required for sensing, processing,
storing, and retrieving information. The majority of sensing frameworks are built based
on modern software design patterns to ensure robustness, security, extendibility, and
openness. The latter two characteristics are highly correlated through the common
ground of allowing third parties to develop their own applications upon these frame-
works but also contribute custom modules to extend and improve them. In addition,
selecting a sensing framework is ostensibly a complicated process. But the designer
should understand that deciding on a certain framework will only constitute a (less)
significant enabler in the application and will not limit its capabilities. Last but not
least, the intelligence that some frameworks provide regarding energy efficiency may
prove to be an additional succor.

Detecting social interactions through mobile devices is a topic that has drawn re-
searchers’ attention. Several approaches have been proposed by leveraging COTS mo-
BILE phones. Researchers have focused on detecting social interactions by utilizing a
single or multiple modalities. Each method has its own advantages and disadvantages.
Most works have performed simple discovery due to the pervasive and robust character
of the approach, given a large amount of error. However, this method does not limit the
user on a specific wearing position, with a very low design and development complexity.
It is less sensitive to environmental factors in comparison to other approaches because
of the large spatial range it covers. The biggest disadvantage of the approach is the
large amount of false positives it provides, especially in crowded places. Researchers
tried to tackle this error through distance estimation based on Bluetooth, WiFi, or au-
dio signals. These approaches are highly dependent on the environment, and human
body absorption constitutes a significant obstacle. Voice and conversation detection
have been incorporated in social interaction inference to increase accuracy. It should
be noted that conversation detection constitutes a great challenge that requires tack-
ing. It is also highly dependent on the environment and the on-body position of the
device. Thus, depending on the accuracy required by the social interaction detection
system, a less or more complex approach could be utilized while also considering the
development effort for each methodology.

In the development of custom mobile devices, the designer decides about the com-
ponents required based on the targeting application. For that reason, there is an ad-
Vantage to selecting a robust and accurate solution (sensor) that will constitute the
appropriate denouement. Although this approach may provide a reliable and robust
solution, the designer must put a lot of effort into limiting its intrusiveness. To this
point, there is no robust and reliable off-the-shelf solution for detecting social interac-
tions on mobile phones in real-world environments.

At the moment, research has focused mainly on extracting behavioral cues because
of immediacy among the device and the cue. This stage is based on the engineering
part of Mobile SSP and does not necessarily require collaboration with psychologists.
Researchers have been mining several types of behavioral cues. Among them are the
auditory for which, although a lot of research was conducted already from SSP, re-
searchers applied some of these techniques on mobile devices. Although the majority
of them were successfully adjusted to smartphones, there are some techniques that increase the computational burden and the energy consumption, and thus, this should be taken into consideration. Physical activity detection is also a topic that has gained researchers’ attention from the point of accelerometer incorporation on COTS mobile phones. As described, this is not a burdensome process and can be executed on mobile phones with high accuracy. Gesture recognition is an arguable cue that to this point required either complex video processing or the user to hold the mobile phone in the hand. This raises questions regarding its real-world applicability. Posture detection is also mainly based on accelerometer data, with(out) external hardware, which can reliably be inferred. However, the on-body position of the device should be included in the process. Facial cues extraction is based primarily on burdensome video processing and object identification. This may not be ideal for continuous sensing applications despite the psychological importance of the cues. Environmental cues provide a significant view of the behavioral cues regarding the context. A reliable and robust solution for detecting interpersonal distance and spatial arrangement from COTS mobile phones in real-world environments is still not available. The device-usage-based cues may not provide information about face-to-face interaction, but it constitutes a lightweight and unobtrusive method that can indicate reliable contextual knowledge. Physiological cues have been extracted through specific external sensors that limit the ubiquitous character of the area. Nevertheless, the type of cues they detect convey information with high significance.

As the literature has indicated, as opposed to extracting behavioral cues, mining social signals and social behaviors on mobile phones is still immature. This occurs for some important reasons that researchers need to take into consideration. Social signals and social behaviors include a noticeable amount of psychological knowledge. They require systematic collaboration with psychologists, which will indicate behavioral cues postulated for mining a certain type of social behavior. To our knowledge, there has not been any tutorial providing a clear guideline of state-of-the-art techniques utilized in each of the steps detecting social signals. A tutorial will provide a definite understanding of the area and the methodology of mining social signals and social behavior. A popular social behavior is stress, which can be detected robustly through auditory, physiological, and physical activity cues. Emotion and mood detection was also mainly performed by auditory, facial, and physiological cues with over 70% accuracy. Different personality traits were primarily detected by auditory, proximity, and phone usage cues, indicating the need for incorporating additional cues in the inference process. Finally, dominance and social role of a person were focused on auditory cues as in SSP, neglecting information such as spatial arrangement.

Overall, Mobile SSP is a multidisciplinary area that requires a considerable amount of knowledge from adjacent fields, indicating the importance of active collaboration. These will drive researchers to incorporate multiple modalities in each of the inference stages. Each of these modalities introduces a certain level of error, intrusiveness, computational burden, and energy consumption that should be considered, as the area targets mobile phones characterized by autonomy issues.

9. CHALLENGES

In the previous section, a discussion about the overall area of Mobile SSP and its main components was presented, identifying the key outcomes of the literature review. This research drove us to conclude, in our opinion, some of the most significant challenges of the area that require tackling. These challenges constitute potential opportunities for research regarding the overall area of Mobile SSP that will provide significant strides in the development and evolution of the area. In the following subsections, each of the
challenges is described, while in some cases initial steps are outlined in order to fill these gaps and to provide a further reference to the reader.

9.1. Context Recognition
Context is one of the most important factors in affective [Zeng et al. 2009] and context-aware computing [Chen and Kotz 2000; Baldauf et al. 2007], anticipatory sensing [Pejovic and Musolesi 2013], and Mobile SSP. As described earlier, SSP delves to interpret social behavior, which requires detection of interactions among people, intertwined with the context in which it is taking place. Acquiring the knowledge of context in a more efficacious way of monitoring and understanding social behavior is looming. Due to the broad meaning of the term, one proposed solution for context recognition is to limit the scope of an application in order to focus on certain aspects of a specific context (e.g., monitor productivity in organizations [Olguín et al. 2009]). However, comprehending and construing context is a great challenge, which requires attentive and systematic research to depict a more holistic view. An example of context recognition is to detect accurately social interactions among people, which will function as a significant enabler of social signal recognition through mobile phones. An important step to understand context is to combine different modalities in a seamless manner to infer social behavior.

9.2. Multimodal Fusion
At this point in time, research has mainly been focusing on extracting various behavioral cues by utilizing different modalities. A limited part of them has tried to infer social behavior, either through individual or by combining a few behavioral cues in a simplistic manner. Due to their continuously increasing computational power, mobile devices allow incessant sensing of various modalities without compromising the user's experience. In order to infer accurately social behavior, merging information from physical and virtual sensors is an indispensable need. Novel fusion techniques may be developed to perform this data amalgamation, precluding information redundancy, increasing the classification accuracy, and mining contingent additional social signals. Targeting the incorporation of multiple modalities through novel fusion techniques, researchers must be able to model the area with the help of psychology to understand which combination of modalities will lead in the identification of certain social behaviors.

9.3. Interdisciplinary Area
Mobile SSP is an area that requires coordination of different fields in each of the stages for mining social behavior. Starting from the sensing layer, experts in different modalities need to cooperate to leverage the most from every modality by providing appropriate preprocessing, fusion, and postprocessing mechanisms. These stages include expertise mainly from electrical and computer engineering such as signal processing. A rife approach to extract behavioral cues and social signals is by utilizing machine-learning techniques. Understanding the type of modalities required to extract a certain form of social behavior indicates that the most important collaboration is between engineering and psychology [Hekler et al. 2013]. Psychologists have great experience in social behavior and could provide guidelines on how to infer different aspects of human behavior. This will supply them with an automatic and concise way to monitor and understand social behavior. In addition, a common challenge among the areas is the issue of acquiring the knowledge of ground truth.

9.4. Ground Truth
Another important challenge in Mobile SSP is establishing ground truth opportunistically in real-world experiments. In state-of-the-art methods including Mobile SSP and
SSP, scientists have acquired ground truth through human observers, camera recordings, or user data labeling. As mentioned before, all three methods are time consuming and prone to human error. Establishing ground truth by asking the user to label the data induces subjectivity from the user’s perspective and eliminates the opportunistic character that is a core idea of the field. It relies on the user’s willingness to provide the experiment’s baseline. Another approach adopted by researchers is to perform experiments in a small-scale and controlled environment such as a room, in order to estimate the accuracy while understanding the method’s limitations. Knowing the limitations of the approach and achieving an acceptable accuracy for a particular application leads to a concrete solution. This enabler is then deployed in a large-scale environment to extract higher-level knowledge of a population with the accuracy that was established in the initial experiments. Although this method has been evaluated in a controlled environment and achieved a particular accuracy, scaling the approach will introduce new sources of error that may need to be tackled. An alternative approach that will be utilized, potentially as Mobile SSP is evolving, is considering as ground truth the outcome of state-of-the-art techniques. However, this method limits an enabler’s results to the state-of-the-art technique’s accuracy. Thus, providing a viable methodology for establishing ground truth in social sciences and especially Mobile SSP while preserving the user’s privacy is an imperative need.

9.5. Privacy
Every application that is directly or indirectly related to humans is also correlated to privacy. For that reason, a very important tradeoff to be made during the design and implementation of a Mobile SSP application is usability against privacy [Avancha et al. 2012]. Regarding usability, in this context, we consider the opportunistic and nonintrusive character of Mobile SSP. The target of a Mobile SSP application is to extract a certain type of behavioral information from the user. However, this target should be achieved with respect to the user’s privacy. Some solutions have been proposed to minimize the impact on the user’s privacy in crowd-sensing applications, where the data are first anonymized and then retrieved from the device [Cornelius et al. 2008]. Privacy could be preserved by performing sensing and inference of social behavior on the user’s device. Thus, the collected data are not transmitted to a third-party application, while the user has the ability to delete unwanted or sensitive information. In some cases, online inference is not applicable due to device resource limitations. In that sense, the designer should introduce a privacy-preserving mechanism that protects users’ anonymity but also allows them to manage and expose only the desirable information in an energy-efficient manner.

9.6. Energy Efficiency
Today’s mobile devices have evolved significantly in terms of sensing and computation during the last decade. But a remaining issue that is challenging researchers in the field of Mobile SSP is battery consumption. To tackle this challenge, scientists may adopt alternative techniques to continuous sensing and inference. One promising approach is to apply adaptive mechanisms (e.g., reinforcement learning) in both sensing and inference regarding the context in which the user/device is in. Another proposed solution is to perform the computations with subtlety either on the device or on a back-end server, in an adjusted manner based on the user’s preference and the device’s status. In order to allow devices to cope with the continuous computational and energetic demand, applications should be able to adapt based on the user’s context, for example, to apply a conservative policy in situations were users’ social behavior is insignificant. Regardless of the existing solutions, we have identified that there is a
great deal of research that has yet to be conducted and requires exploitation in each of the stages during the inference process.

10. CONCLUSIONS

After Pentland’s introduction of Honest Signals [Pentland 2008], the research community focused on modeling, analyzing, and synthesizing human behavior in an automatic manner. This interest was raised mainly due to the novel point of view introduced by incorporating the social, spontaneous, and native aspect of human behavior. Capturing this type of physical signals is a challenge, but mobile devices with the pervasive, ubiquitous, and unobtrusive characteristics are a candidate solution. Mobile devices are personalized tools that are able through intelligent learning techniques to adopt to their users’ preferences. Additionally, they eliminate the person detection process of SSP and thus provide more accurate results through less computationally demanding processes. In our opinion, Mobile SSP is a promising area but requires a great deal of effort to overcome its main challenges. The scientific community has to finalize the core term-definition in order to establish a common ground. There have been noticeable works at lower layers of extracting social behavior on mobile devices, for example, open-source sensing and context recognition frameworks that provide an important abstraction enhancement. Currently, to our knowledge, there is no concrete framework for detecting and measuring social interactions on mobile phones in contrast to wearable devices that are able to accurately identify face-to-face interactions. Also, context-recognition-based works have to be leveraged and combined with the theoretical knowledge from the field of psychology. This will lead to modeling and analyzing an additional sizeable amount of various behavioral cues in an energy-efficient way. However, mining social signals and combining them to infer a user’s social behavior is still an area in which limited research has been conducted due to lack of coordination with the field of psychology. By tackling the challenges of Mobile SSP, a new realm will emerge with applications in several fields and provide numerous benefits to areas such as healthcare, organizational engineering, and marketing.

REFERENCES

G. Adomavicius and A. Tuzhilin. 2011. Context-aware recommender systems. In Recommender Systems Handbook, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor (Eds.). Springer US, 217–253.

N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland. 2011. Social fMRI: Investigating and shaping social mechanisms in the real world. Pervasive Mob. Comput. 7, 6 (Dec. 2011), 643–659.

Y. Altshuler, M. Fire, N. Aharony, Z. Volkovich, Y. Elovici, and A. Pentland. 2013. Trade-offs in social and behavioral modeling in mobile networks. In Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction (SBP’13). Springer-Verlag, Heidelberg, 412–423.

A. K. Amin, B. Kersten, O. A. Kulyk, E. Pelgrim, J. Wang, and P. Markopoulos. 2005. The SenseMS: Enriching the SMS experience for teens by non-verbal means. In Human-Computer Interaction (INTERACT’05), Maria Francesca Costabile and Fabio Patern (Eds.). Lecture Notes in Computer Science, Vol. 3585. Springer, Heidelberg, 962–965.

Y. Anokwa, C. Hartung, W. Brunette, G. Borriello, and A. Lerer. 2009. Open source data collection in the developing world. Computer 42, 10 (Oct. 2009), 97–99.

A. Antoniou, E. Theodoridis, I. Chatzigiannakis, and G. Mylonas. 2011. Monitoring physical space using mobile phones for inferring social and contextual interactions. In Proceedings of the 2011 IEEE Sensors. 1616–1619.

O. Aran and D. Gatica-Perez. 2011. Analysis of group conversations: Modeling social verticality. In Computer Analysis of Human Behavior, Albert Ali Salah and Theo Gevers (Eds.). Springer, London, 293–322.

S. Avancha, A. Baxi, and D. Kotz. 2012. Privacy in mobile technology for personal healthcare. ACM Comput. Surv. 45, 1, Article 3 (Dec. 2012), 54 pages.
M. Azizyan, I. Constandache, and R. Roy Choudhury. 2009. SurroundSense: Mobile phone localization via ambience fingerprinting. In Proceedings of the 15th Annual International Conference on Mobile Computing and Networking (MobiCom’09). ACM, New York, NY, 261–272.

M. Baldauf, S. Dustdar, and F. Rosenberg. 2007. A survey on context-aware systems. Int. J. Ad Hoc Ubiquitous Comput. 2, 4 (June 2007), 263–277.

N. Banerjee, S. Agarwal, P. Bahl, R. Chandra, A. Wolman, and M. Corner. 2010. Virtual compass: Relative positioning to sense mobile social interactions. In Proceedings of the 6th International Conference on Pervasive Computing (Pervasive’10). Springer-Verlag, Berlin, 1–21.

M. R. Barrich and M. K. Mount. 1991. The big five personality dimensions and job performance: A meta analysis. Personnel Psychol. 44, 1 (1991), 1–26.

S. Basu. 2003. A linked-HMM model for robust voicing and speech detection. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003, Proceedings (ICASSP'03), Vol. 1. 1–816–1–819.

T. S. Bateman and D. W. Organ. 1983. Job satisfaction and the good soldier: The relationship between affect and employee citizenship. Acad. Management J. 26, 4 (1983), 587–595.

G. Bauer and P. Lukowicz. 2012. Can smartphones detect stress-related changes in the behaviour of individuals? In Proceedings of the 2012 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops). 423–426.

P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell. 19, 7 (July 1997), 711–720.

E. M. Berke, T. Choudhury, S. Ali, and M. Rabbi. 2011. Objective measurement of sociability and activity: Mobile sensing in the community. Ann. Fam. Med. 9, 4 (2011), 344–350.

A. Bogomolov, B. Lepri, and F. Pianesi. 2013. Happiness recognition from mobile phone data. In Proceedings of the 2013 International Conference on Social Computing (SocialCom’13), 790–795.

G. R. Bradski. 1998. Computer Vision Face Tracking For Use in a Perceptual User Interface. (1998).

P. M. Brunet, R. Cowie, D. Heylen, A. Nijholt, and M. Schröder. 2012. Conceptual frameworks for multimodal social signal processing. J. Multimodal User Interfaces 6, 3–4 (2012), 95–99.

A. Bulling, U. Blanke, and B. Schiele. 2014. A tutorial on human activity recognition using body-worn inertial sensors. ACM Comput. Surv. 46, 3, Article 33 (Jan. 2014), 33 pages.

A. Bulling, D. Roggen, and G. Tröster. 2011. What’s in the eyes for context-awareness? IEEE Pervasive Comput. 10, 2 (Apr. 2011), 48–57.

A. Campbell, T. Choudhury, S. Hu, H. Lu, M. K. Mukerjee, M. Rabbi, and R. D. S. Raizada. 2010. NeuroPhone: Brain-mobile phone interface using a wireless EEG headset. In Proceedings of the 2nd ACM SIGCOMM Workshop on Networking, Systems, and Applications on Mobile Handhelds (MobiHeld’10). ACM, New York, NY, 3–8.

N. Carbonaro, G. Anania, G. D. Mura, M. Tesconi, A. Tognetti, G. Zupone, and D. De Rossi. 2011. Wearable biomonitoring system for stress management: A preliminary study on robust ECG signal processing. In Proceedings of the 2011 IEEE Symposium on a World of Wireless, Mobile and Multimodal Networks (WOWMOM’11). IEEE Computer Society, Washington, DC, 1–6.

G. Cardone, A. Cirri, A. Corradi, L. Foschini, and D. Maio. 2013. MSF: An efficient mobile phone sensing framework. Int. J. Distrib. Sensor Netw. 2013, 538937 (2013), 9.

D. Carlson and A. Schrader. 2012. Dynamix: An open plug-and-play context framework for android. In Proceedings of the 2012 3rd International Conference on the Internet of Things (IOT). 151–158.

I. Carreras, A. Matei, P. Saar, and V. Osmani. 2012. Comm2Sense: Detecting proximity through smartphones. In Proceedings of the 2012 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops). 253–258.

K. Chang, D. Fisher, J. Canny, and B. Hartmann. 2011. How’s my mood and stress? An efficient speech analysis library for noninvasive monitoring on mobile phones. In Proceedings of the 6th International Conference on Body Area Networks (BodyNets’11). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, 71–77.

T. L. Chartrand and J. A. Bargh. 1999. The chameleon effect: The perception-behavior link and social interaction. J. Pers. Social Psychol. 76, 6 (1999), 893–910.

G. Chen and D. Kotz. 2000. A Survey of Context-Aware Mobile Computing Research. Technical Report. Hanover, NH, USA.

Y. Chen, D. Pavlov, and J. F. Canny. 2009. Large-scale behavioral targeting. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’09). ACM, New York, NY, 209–218.
A Survey on Mobile Social Signal Processing

H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin. 2010. Diversity in smartphone usage. In Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services (MobiSys'10). ACM, New York, NY, 179–194.

S. Feese, B. Arnrich, G. Troster, M. Burtscher, B. Meyer, and K. Jonas. 2013. CoenoFire: Monitoring performance indicators of firefighters in real-world missions using smartphones. In Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp’13). ACM, New York, NY, 83–92.

V. Filonenko, C. Cullen, and J. Carswell. 2010. Investigating ultrasonic positioning on mobile phones. In Proceedings of the 2010 International Conference on Indoor Positioning and Indoor Navigation (IPIN’10). 1–8.

V. Filonenko, C. Cullen, and J. D. Carswell. 2013. Indoor positioning for smartphones using asynchronous ultrasound trilateration. ISPRS Int. J. Geo-Information 2, 3 (2013), 589–620.

Y. Freund and R. E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 1 (Aug. 1997), 119–139.

N. Friedman, D. Geiger, and M. Goldszmidt. 1997. Bayesian network classifiers. Mach. Learn. 29, 2–3 (Nov. 1997), 131–163.

J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay. 2007. MyExperience: A system for in situ tracing and capturing of user feedback on mobile phones. In Proceedings of the 5th International Conference on Mobile Systems, Applications and Services (MobiSys’07). ACM, New York, NY, 57–70.

Funf. 2011. Open Sensing Framework. (2011). http://www.funf.org.

A. Gaggioli, G. Pioggia, G. Tartarisco, G. Baldus, M. Ferro, P. Cipresso, S. Serino, A. Popleteev, S. Gabrielli, R. Maimone, and G. Riva. 2012. A system for automatic detection of momentary stress in naturalistic settings. Studies in Health Technol. Informatics 181 (2012), 182–186.

M. P. Gardner. 1985. Mood states and consumer behavior: A critical review. J. Consumer Res. 12, 3 (1985), pp. 281–300.

A. Ghose, C. Bhaumik, and T. Chakravarty. 2013. BlueEye: A system for proximity detection using bluetooth on mobile phones. In Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication (UbiComp’13 Adjunct). ACM, New York, NY, 1135–1142.

P. K. Ghosh, A. Tsiartas, and S. Narayanan. 2011. Robust voice activity detection using long-term signal variability. IEEE Trans. Audio, Speech, and Language Proc. 19, 3 (March 2011), 600–613.

A. Gluhak, M. Presser, L. Zhu, S. Esfandiyari, and S. Kupschick. 2007. Towards mood based mobile services and applications. In Proceedings of the 2nd European Conference on Smart Sensing and Context (EuroSSC’07). Springer-Verlag, Berlin, 159–174.

E. Goffman. 1963. Behavior in Public Places: Notes on the Social Organization of Gatherings. Free Press.

L. R. Goldberg. 1993. The structure of phenotypic personality traits. Am. Psychologist 48, 1 (1993), 26–34.

Google Glass. 2013. Wearable Technology. (2013). http://www.google.com/glass.

M. Gorgoglione, U. Panniello, and A. Tuzhilin. 2011. The effect of context-aware recommendations on customer purchasing behavior and trust. In Proceedings of the 5th ACM Conference on Recommender Systems (RecSys’11). ACM, New York, NY, 85–92.

G. Groh, A. Lehmann, J. Reimers, M. R. Friess, and L. Schwarz. 2010. Detecting social situations from interaction geometry. In Proceedings of the 2010 IEEE Second International Conference on Social Computing (SOCIALCOM’10). IEEE Computer Society, Washington, DC, 1–8.

R. M. Groves. 2004. Survey Errors and Survey Costs. Vol. 536. John Wiley & Sons.

E. T. Hall. 1966. The Hidden Dimension. Doubleday.

M. A. Hall and G. Holmes. 2003. Benchmarking attribute selection techniques for discrete class data mining. IEEE Trans. Knowl. Data Eng. 15, 6 (Nov. 2003), 1437–1447.

E. B. Hekler, P. Klasnja, J. E. Froehlich, and M. P. Buman. 2013. Mind the theoretical gap: Interpreting, using, and developing behavioral theory in HCI research. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI’13). ACM, New York, NY, 3307–3316.

E. Hjelmas and B. K. Low. 2001. Face detection: A survey. Comput. Vis. Image Understanding 83, 3 (2001), 236–274.

D. Hong, B. Zhang, Q. Li, S. Nirjon, R. Dickerson, G. Shen, X. Jiang, and J. Stankovic. 2012. SEPTIMU: Continuous in-situ human wellness monitoring and feedback using sensors embedded in earphones. In Proceedings of the 11th International Conference on Information Processing in Sensor Networks (IPSN’12). ACM, New York, NY, 159–160.

S. A. Hoseininatababaei, A. Gluhak, and R. Tafazolli. 2013. A survey on smartphone-based systems for opportunistic user context recognition. ACM Comput. Surv. 45, 3, Article 27 (July 2013), 51 pages.
S. A. Hoseinitabatabaei, A. Gluhak, and R. Tafazolli. 2014. Towards a position and orientation independent approach for pervasive observation of user direction with mobile phones. *Pervasive Mobile Comput.* 0 (2014), 23–42.

S. A. Hoseinitabatabaei, A. Gluhak, R. Tafazolli, and W. Headley. 2013. Design, realization, and evaluation of uDirect – An approach for pervasive observation of user facing direction on mobile phones. *IEEE Trans. Mobile Comput.* 99, PrePrints (2013), 1.

W. Hu, G. Cao, S. V. Krishnamurthy, and P. Mohapatra. 2013. Mobility-assisted energy-aware user contact detection in mobile social networks. In *Proceedings of the 2013 IEEE 33rd International Conference on Distributed Computing Systems*. 155–164.

G. Hua, T. Y. Yang, and S. Vasiредdy. 2007. PEYE: Toward a visual motion based perceptual interface for mobile devices. In *Proceedings of the 2007 IEEE International Conference on Human-computer Interaction (HCI’07)*. Springer-Verlag, Berlin, 39–48.

S. Hwang and K. Wohn. 2013. VibroTactor: Low-cost placement-aware technique using vibration echoes on mobile devices. In *Proceedings of the Companion Publication of the 2013 International Conference on Intelligent User Interfaces Companion (IUI’13 Companion)*. ACM, New York, NY, 73–74.

A. Iamnitchi, J. Blackburn, and N. Kourtellis. 2012. The social hourglass: An infrastructure for socially aware applications and services. *IEEE Internet Comput.* 16, 3 (May 2012), 15–23.

F. Ichikawa, J. Chipchase, and R. Grignani. 2005. Where’s the phone? A study of mobile phone location in public spaces. In *Proceedings of the 2005 2nd International Conference on Mobile Technology, Applications and Systems*. 1–8.

G. H. John and P. Langley. 1995. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI’95)*. Morgan Kaufmann Publishers, San Francisco, CA, 338–345.

E. Jovanov, M. Milosevic, and A. Milenkovic. 2013. A mobile system for assessment of physiological response to posture transitions. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC’13)*. 7205–7208.

H. Junker, O. Amft, P. Lukowicz, and G. Troster. 2008. Gesture spotting with body-worn inertial sensors to detect user activities. *Pattern Recogn.* 41, 6 (2008), 2010–2024.

S. Kang, J. Lee, H. Jang, H. Lee, Y. Lee, S. Park, T. Park, and J. Song. 2008. SeeMon: Scalable and energy-efficient context monitoring framework for sensor-rich mobile environments. In *Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services (MobiSys’08)*. ACM, New York, NY, 267–280.

M. Keally, G. Zhou, G. Xing, J. Wu, and A. Pyles. 2011. PBN: Towards practical activity recognition using smartphone-based body sensor networks. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems (SenSys’11)*. ACM, New York, NY, 246–259.

A. Kendon. 1990. *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge University Press.

E. J. Keogh, S. Chu, D. Hart, and M. J. Pazzani. 2001. An online algorithm for segmenting time series. In *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM’01)*. IEEE Computer Society, Washington, DC, 289–296.

T. Kim, A. Chang, L. Holland, and A. Pentland. 2008. Meeting mediator: Enhancing group collaboration using sociometric feedback. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW’08)*. ACM, New York, NY, 457–466.

N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. 2010. A survey of mobile phone sensing. *Comm. Mag.* 48, 9 (Sept. 2010), 49–56.
A Survey on Mobile Social Signal Processing

of the 13th International Conference on Ubiquitous Computing (UbiComp’11). ACM, New York, NY, 355–364.

H. Lee, Y. S. Choi, and S. Lee. 2012. Mobile posture monitoring system to prevent physical health risk of smartphone users. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp’12). ACM, New York, NY, 592–593.

H. Lee, Y. S. Choi, S. Lee, and E. Shim. 2013a. Smart pose: Mobile posture-aware system for lowering physical health risk of smartphone users. In CHI’13 Extended Abstracts on Human Factors in Computing Systems (CHI EA’13). ACM, New York, NY, 2257–2266.

Y. Lee, C. Min, C. Hwang, J. Lee, I. Hwang, Y. Ju, C. Yoo, M. Moon, U. Lee, and J. Song. 2013b. SocioPhone: Everyday face-to-face interaction monitoring platform using multi-phone sensor fusion. In Proceedings of the 11th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys’13). ACM, New York, NY, 375–388.

D. Li, D. Winfield, and D. J. Parkhurst. 2005. Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05) - Workshops - Volume 03 (CVPR’05). IEEE Computer Society, Washington, DC, 79.

K. A. Li, T. Y. Sohn, S. Huang, and W. G. Griswold. 2008. Peopletones: A system for the detection and notification of buddy proximity on mobile phones. In Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services (MobiSys’08). ACM, New York, NY, 160–173.

Q. Li, S. Chen, and J. A. Stankovic. 2013. Multi-modal in-person interaction monitoring using smartphone and on-body sensors. In 2013 IEEE International Conference on Body Sensor Networks (BSN’13). 1–6.

M. Liberman, K. Davis, M. Grossman, N. Martey, and J. Bell. 2002. Emotional Prosody Speech and Transcripts. Linguistic Data Consortium, Philadelphia.

R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong. 2013. MoodScope: Building a mood sensor from smartphone usage patterns. In Proceedings of the 11th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys’13). ACM, New York, NY, 389–402.

J. Liu, C. Wang, J. Qiao, W. Wang, and Y. Zhang. 2012. A machine-to-machine application: Body posture recognition on smartphones for intelligent remote control. In Proceedings of the 2012 IEEE 2nd International Conference on Cloud Computing and Intelligent Systems (CCIS’12), Vol. 02. 884–888.

J. Liu, L. Zhong, J. Wickramasuriya, and V. Vasudevan. 2009. uWave: Accelerometer-based personalized gesture recognition and its applications. Pervasive Mob. Comput. 5, 6 (Dec. 2009), 657–675.

K. Liu, X. Liu, and X. Li. 2013b. Guoguo: Enabling fine-grained indoor localization via smartphone. In Proceedings of the 11th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys’13). ACM, New York, NY, 235–248.

K. Liu, X. Liu, L. Xie, and X. Li. 2013c. Towards accurate acoustic localization on a smartphone. In Proceedings of the 2013 IEEE INFOCOM. 495–499.

S. Liu, Y. Jiang, and A. Striegel. 2013a. Face-to-face proximity estimation using bluetooth on smartphones. IEEE Trans. Mob. Comput. 99, PrePrints (2013), 1.

H. Lu, A. J. B. Brush, B. Priyantha, A. K. Karlson, and J. Liu. 2011. SpeakerSense: Energy efficient unobtrusive speaker identification on mobile phones. In Proceedings of the 9th International Conference on Pervasive Computing (Pervasive’11). Springer-Verlag, Berlin, 188–205.

H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury. 2012. StressSense: Detecting stress in unconstrained acoustic environments using smartphones. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp’12). ACM, New York, NY, 351–360.

H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell. 2009. SoundSense: Scalable sound sensing for people-centric applications on mobile phones. In Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services (MobiSys’09). ACM, New York, NY, 165–178.

H. Lu, J. Yang, Z. Liu, N. D. Lane, T. Choudhury, and A. T. Campbell. 2010. The jigsaw continuous sensing engine for mobile phone applications. In Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems (SenSys’10). ACM, New York, NY, 71–84.

M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. 1998. Coding facial expressions with Gabor wavelets. In Proceedings of the 3rd International Conference on Face & Gesture Recognition (FG’98). IEEE Computer Society, Washington, DC, 200.

A. Madan and A. Pentland. 2006. VibeFones: Socially aware mobile phones. In Proceedings of the 10th IEEE International Symposium on Wearable Computers. 2006. 109–112.

A. Matic, V. Osmani, A. Maxhuni, and O. Mayora. 2012. Multi-modal mobile sensing of social interactions. In Proceedings of the 2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth). 105–114.
I. Matthews and S. Baker. 2004. Active appearance models revisited. *Int. J. Comput. Vision* 60, 2 (Nov. 2004), 135–164.

E. Miluzzo, C. T. Cornelius, A. Ramaswamy, T. Choudhury, Z. Liu, and A. T. Campbell. 2010. Darwin phones: The evolution of sensing and inference on mobile phones. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services (MobiSys’10)*. ACM, New York, NY, 5–20.

E. Miluzzo, N. D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. B. Eisenman, X. Zheng, and A. T. Campbell. 2008. Sensing meets mobile social networks: The design, implementation and evaluation of the CenceMe application. In *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems (SenSys’08)*. ACM, New York, NY, 337–350.

S. Mitra and T. Acharya. 2007. Gesture recognition: A survey. *IEEE Trans. Syst. Man Cybernet. Part C: Appl. Rev.* 37, 3 (May 2007), 311–324.

A. Muaremi, B. Arnrich, and G. Trster. 2013. Towards measuring stress with smartphones and wearable devices during workday and sleep. *BioNanoScience* 3, 2 (2013), 172–183.

Myo. 2013. Gesture Control Armband. Retrieved from https://www.thalmic.com/en/myo/.

E. Miluzzo, N. D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. B. Eisenman, X. Zheng, and A. T. Campbell. 2008. Sensing meets mobile social networks: The design, implementation and evaluation of the CenceMe application. In *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems (SenSys’08)*. ACM, New York, NY, 337–350.

S. Mitra and T. Acharya. 2007. Gesture recognition: A survey. *IEEE Trans. Syst. Man Cybernet. Part C: Appl. Rev.* 37, 3 (May 2007), 311–324.

A. Muaremi, B. Arnrich, and G. Trster. 2013. Towards measuring stress with smartphones and wearable devices during workday and sleep. *BioNanoScience* 3, 2 (2013), 172–183.

Myo. 2013. Gesture Control Armband. Retrieved from https://www.thalmic.com/en/myo/.

S. Nirjon, R. F. Dickerson, P. Asare, Q. Li, D. Hong, J. A. Stankovic, P. Hu, G. Shen, and X. Jiang. 2013. Auditeur: A mobile-cloud service platform for acoustic event detection on smartphones. In *Proceedings of the 11th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys’13)*. ACM, New York, NY, 403–416.

D. Olguín, B. N. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland. 2009. Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *Trans. Syst. Man Cyber Part B 39, 1* (Feb. 2009), 43–55.

E. Oliver. 2010. The challenges in large-scale smartphone user studies. In *Proceedings of the 2nd ACM International Workshop on Hot Topics in Planet-scale Measurement (HotPlanet’10)*. ACM, New York, NY, Article 5, 5 pages.

V. Osmani, I. Carreras, A. Matic, and P. Saar. 2014. An analysis of distance estimation to detect proximity in social interactions. *J. Ambient Intell. Human. Comput.* 5, 3 (2014), 297–306.

A. Pantelopoulos and N. G. Bourbakis. 2010. A survey on wearable-sensor-based systems for health monitoring and prognosis. *Trans. Sys. Man Cyber Part C 40, 1* (Jan. 2010), 1–12.

T. Park, J. Lee, I. Hwang, C. Yoo, L. Nachman, and J. Song. 2011. E-gesture: A collaborative architecture for energy-efficient gesture recognition with hand-worn sensor and mobile devices. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems (SenSys’11)*. ACM, New York, NY, 260–273.

V. Pejovic and M. Musolesi. 2013. Anticipatory mobile computing: A survey of the state of the art and research challenges. *CoRR* abs/1306.2356 (2013).

C. Peng, G. Shen, Y. Zhang, Y. Li, and K. Tan. 2007. BeepBeep: A high accuracy acoustic ranging system using COTS mobile devices. In *Proceedings of the 5th International Conference on Embedded Networked Sensor Systems (SenSys’07)*. ACM, New York, NY, 1–14.

A. Pentland. 2007. Social signal processing [exploratory DSP]. *IEEE Signal Process. Mag.* 24, 4 (July 2007), 108–111.

A. Pentland. 2008. *Honest Signals: How They Shape Our World*. The MIT Press.

A. Pentland. 2012. The new science of building great teams. *Harvard Business Rev.* 90, 4 (2012), 60–69.

K. Peternel, M. Poganik, R. Tavar, and A. Kos. 2012. A presence-based context-aware chronic stress recognition system. *Sensors* 12, 11 (2012), 15888–15906.

C. Pino and I. Kavasidis. 2012. Improving mobile device interaction by eye tracking analysis. In *Proceedings of the 2012 Federated Conference on Computer Science and Information Systems (FedCSIS’12)*. 1199–1202.

I. Poggi, F. DErrico, and A. Vinciarelli. 2012. Social signals: From theory to applications. *Cogn. Process.* 13, 2 (2012), 389–396.

R. J. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Francisco, CA, USA.

M. R. Ra, B. Liu, T. La Porta, and R. Govindan. 2012. Medusa: A programming framework for crowd-sensing applications. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services (MobiSys’12)*.

K. K. Rachuri, C. Efstratiou, I. Leontiadis, C. Mascolo, and P. J. Rentfrow. 2013. METIS: Exploring mobile phone sensing offloading for efficiently supporting social sensing applications. In *Proceedings of the 2013 IEEE International Conference on Pervasive Computing and Communications (PerCom’13)*. 85–93.

K. K. Rachuri, C. Mascolo, M. Musolesi, and P. J. Rentfrow. 2011. SociableSense: Exploring the trade-offs of adaptive sampling and computation offloading for social sensing. In *Proceedings of the 17th Annual
A Survey on Mobile Social Signal Processing

International Conference on Mobile Computing and Networking (MobiCom’11). ACM, New York, NY, 73–84.

K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas. 2010. EmotionSense: A mobile phones based adaptive platform for experimental social psychology research. In Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp’10). ACM, New York, NY, 281–290.

A. A. Razak and M. I. Z. Abidin. 2005. Comparison between fuzzy and NN method for speech emotion recognition. In Proceedings of the 3rd International Conference on Information Technology and Applications (ICITA05), Volume 2. IEEE Computer Society, Washington, DC, 297–302.

J. Reason. 1990. Human Error. Cambridge University Press.

A. Sashima, Y. Inoue, T. Ikeda, T. Yamashita, and K. Kurumatani. 2008. CONSORTS-S: A mobile sensing platform for context-aware services. In Proceedings of the International Conference on Intelligent Sensors, Sensor Networks and Information Processing, 2008 (ISSNIP08). 417–422.

J. Saunders. 1996. Real-time discrimination of broadcast speech/music. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996 (ICASSP’96)., Vol. 2. 993–996.

E. A. Schegloff. 2000. Overlapping talk and the organization of turn-taking for conversation. Lang. Society 29, 01 (2000), 1–63.

M. Schrder. 2003. Experimental study of affect bursts. Speech Comm. 40, 12 (2003), 99–116.

J. Seiter, S. Feese, B. Arnrich, G. Tröster, O. Amft, L. Macrea, and K. Maurer. 2013. Activity monitoring in daily life as an outcome measure for surgical pain relief intervention using smartphones. In Proceedings of the 2013 International Symposium on Wearable Computers (ISWC’13). ACM, New York, NY, 127–128.

E. Selkirk and J. A. Goldsmith. 1995. Sentence prosody: Intonation, stress, and phrasing. In The Handbook of Phonological Theory, Vol. 1. 550–569.

C. Shepard, A. Rahmati, C. Tossell, L. Zhong, and P. Kortum. 2011. LiveLab: Measuring wireless networks and smartphone users in the field. SIGMETRICS Perform. Eval. Rev. 38, 3 (Jan. 2011), 15–20.

Y. Shi, Y. Shi, and J. Liu. 2011. A rotation based method for detecting on-body positions of mobile devices. In Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp’11). ACM, New York, NY, 559–560.

V. K. Singh, L. Freeman, B. Lepri, and A. Pentland. 2013a. Predicting spending behavior using socio-mobile features. In Proceedings of the 2013 International Conference on Social Computing (SocialCom’13). 174–179.

V. K. Singh, L. Freeman, B. Lepri, and A. Pentland. 2013b. Predicting spending behavior using socio-mobile features. In Proceedings of the 2010 IEEE International Conference on Social Computing/IEEE International Conference on Privacy, Security, Risk and Trust. 174–179.

V. Srinivasan and T. Phan. 2012. An accurate two-tier classifier for efficient duty-cycling of smartphone activity recognition systems. In Proceedings of the 3rd International Workshop on Sensing Applications on Mobile Phones (PhoneSense’12). ACM, New York, NY, Article 11, 5 pages.

J. Staiano, B. Lepri, N. Aharon, F. Pianesi, N. Sebe, and A. Pentland. 2012. Friends don’t lie: Inferring personality traits from social network structure. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp’12). ACM, New York, NY, 321–330.

J. Stehl, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J. F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Rgis, B. Lina, and P. Vanhems. 2011. High-resolution measurements of face-to-face contact patterns in a primary school. PLOS ONE 6, 8 (2011), e23176.

M. Strohbach and M. Martin. 2011. Toward a platform for pervasive display applications in retail environments. IEEE Pervasive Comput. 10, 2 (2011), 19–27.

F. T. Sun, C. Kuo, H. T. Cheng, S. Buthpitiya, P. Collins, and M. Griss. 2012. Activity-aware mental stress detection using physiological sensors. In Mobile Computing, Applications, and Services, Martin Gris and Guang Yang (Eds.). Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Vol. 76. Springer, Berlin, 211–230.

P. Sundström, A. Ståhl, and K. Höök. 2005. eMoto: Affectively involving both body and mind. In CHI’05 Extended Abstracts on Human Factors in Computing Systems (CHI EA05). ACM, New York, NY, 2005–2008.

TagMobile. 2003. RFID Centre: Indoor RFID Tracking. Retrieved from http://www.ubisense.net/en/index.html.

M. Turk and A. Pentland. 1991. Eigenfaces for recognition. J. Cogn. Neurosci. 3, 1 (Jan. 1991), 71–86.

Ubisense. 2003. Location Intelligence solutions. Retrieved from http://www.ubisense.net/en/index.html.

M. Uddin and T. Nadeem. 2013. RF-beep: A light ranging scheme for smart devices. In 2013 IEEE International Conference on Pervasive Computing and Communications (PerCom’13). 114–122.
