Crowdsourcing solutions for data gathering from wearables

Lucie Klus[(1)], Elena Simona Lohan[(1)], Carlos Granell[(2)], Jari Nurmi[(1)]

[(1)] Tampere University, Tampere, Finland, name.surname@tuni.fi

[(2)] Universitat Jaume I, Castellon, Spain, carlos.granell@uji.es

Abstract

This paper gives an overview of crowdsourcing databases and crowdsourcing-related challenges and open research issues for data collected from wearable devices. It is shown that, with the advent of smarter wearable devices, the complexity of data gathering, storage, and processing in crowdsourced modes will increase exponentially and new solutions are needed in order to cope with larger data sets and low energy consumption in wearable devices, while ensuring the integrity and quality of the collected data.

1 Introduction

Wearable devices, as functional and fashionable accessories, have been around for centuries, although their popularity, appearance, and use has evolved over time. From the first pedometer invented by Leonardo da Vinci in 15th century, through wristwatches, hearing aids, or heart rate monitors, the technological advancements today allow us to measure and gather large amounts of information from a single, small, and user-friendly device. Gathering information from a large number of users enables new, until-recently-undiscovered possibilities of what to do with data and how to learn new things from it. The phenomenon known as *data mining* refers to the process of extracting new useful information from a large data set. In order to "mine" the information and to efficiently store the data, the raw data coming from a wearable device must be pre-processed. This paper presents an overview of the main data types collected from wearables, lists available repositories with wearable-oriented data, and discusses the main questions around crowdsourced wearable data, from data collection and pre-processing to data storage and analysis.

2 Wearables data types

The future smart wearables will most likely include a large variety of devices with different sensors, processing capabilities, and connectivity options. A key aspect will be the type and volume of data they gather and process, which can be divided into two basic groups: sensor-based data and application-based data.

Sensor-based data include information derived from the embedded sensors of the wearable such as physiological data (heart rate, number of steps, body temperature, blood pressure), data about the surroundings (temperature, atmospheric pressure, humidity, air pollution), and location-based data (GPS coordinates, altitude).

Application-based data refer to the data collected from a wearable-based app, e.g. running on a smartphone. Apart from the voluminous entertainment-centered data such as music or video streaming, these data typically include personal data (instant messages, internet profiles, payment information). Data leakage may lead to serious concerns such as property or identity theft.

The literature mentions several data-collection categories depending on the gathering frequency and user's involvement [1]. For the latter, for example, *participatory sensing* requires user's action to capture and share data, while in *opportunistic sensing* the mobile app gathers sensor-related information without user's awareness. Going one step further, opportunistic mobile social networks allow direct interaction, communication and information sharing between mobile devices based on their proximity, without the intervention of people.

3 Open-source and proprietary data repositories

Table 1 summarizes the main datasets for wearables currently existing in the research community as compiled by the authors. The last column (publications) shows either the repository link or gives examples of research papers where the corresponding dataset has been used.

Table 1: Available wearable datasets in literature.

Repository	Format	Data types	Approx. Size	Open-access?	Publications
Insight4wear	.txt	Heart rate, light sensor, battery, application info	500 MB	Yes*	[2]
Zenodo	Many (e.g. .csv)	Many (e.g. GPS, time, orientation, heart rate)	Up to GBs	Yes	[3]
UCI Repository[1]	Many (e.g. .dat)	Many (e.g. accelerometer, age, BMI, GPS)	Up to GBs	Yes	[4]
CrowdSignals	JSON	Many (e.g. location, sensor, network logs)	N/A	No*	[5]
CRAWDAD	Many (e.g. .csv)	Many (e.g. network information, GPS)	Up to GBs	Yes	[6]
PhysioNet	Many (e.g. .csv)	Many (e.g. accelerometers, gyroscopes, ECG)	Up to GBs	Yes	[7]
Kaggle	Many (e.g. .csv)	Many (e.g. compass, human activity data)	Up to GBs	Yes	[8]
Microsoft Research Open Data	Many (e.g. .csv)	Many (e.g. gyroscope, repetitions, accelerometer, heart rate, GPS)	Up to GBs	Yes	[9]
GitHub	Many (e.g. .dat)	Many (e.g. breath and heart rate, GPS)	Up to GBs	Yes	[10]
DRYAD	Many (e.g. .csv)	Many (e.g. temperature, humidity, GPS)	Up to GBs	Yes	[11]
data.gov	JSON	Many (e.g. blood pressure, heart rate, CO_2)	Up to GBs	Yes	[12]

* Available upon request for research purposes or free sample.

4 Open research questions related to crowdsourced data from wearables

As the smart wearable technologies are still on the rise, when it comes to crowdsourcing data from wearables, the available solutions for data collection, storage, processing and transmission are still limited. Next, we identify the main research questions in these four stages, as shown in Fig. 1.

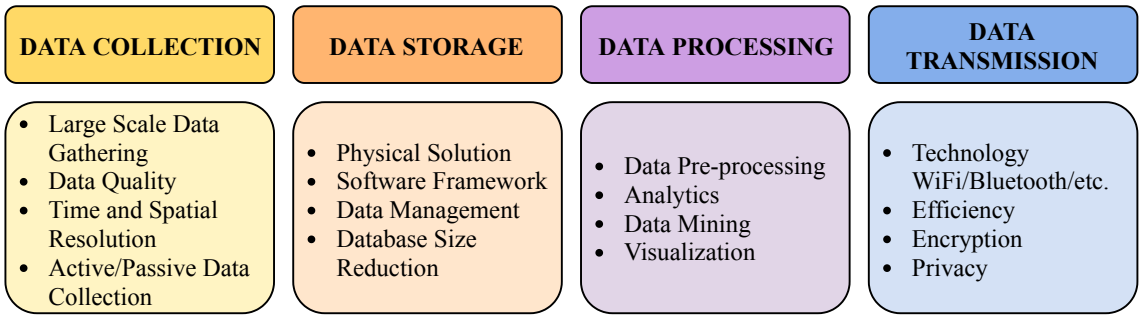

Fig. 1: Summary of main research questions related to crowdsourced data from wearables.

The first stage in Fig. 1 is **data collection**, where research questions pertain to as how to obtain

sufficiently large measurement datasets from wearables in order to enable statistically relevant processing for a variety of wearables-based services and applications. The second stage in Fig. 1 is **data storage**, which refers to how and where the wearable-based data is stored. The main research questions related to the **data processing** stage are connected to how to efficiently process data, preferably in real time, how to obtain new information from collected data, and how to detect and automatically remove outliers. Last but not least, the **data transmission** stage refers to the wireless connectivity between the wearable and the gateway (e.g., mobile device, access node) or directly with the cloud. Here, it is not only necessary to find low-cost low-power connectivity solutions, but it is also important to have an efficient and robust transmission chain. Some existing solutions encountered in the literature with respect to the above-mentioned open challenges are depicted in the bottom row of Fig. 1 and described briefly below.

For the **data collection** stage, to obtain sufficiently large datasets to enable subsequent processing operations over the data, one could use diverse strategies:

- **Dummy datasets** created using a simulation software, which can provide data according to pre-defined types and distributions. This solution allows further work on storage and processing efficiency, but it becomes hard to obtain new information from such data;
- **Merged datasets** refer to creating a single, large dataset from available ones. As data may come in varied types, formats and semantics, data integrate remain a challenge. Therefore, a solid, dependable solution to keep the merged repository operable is required;
- **Proprietary datasets** are mostly obtained from private companies (e.g., Amazon, Apple, Fitbit) or research units owning these types of smart wearables. It is the simplest solution. Yet, in the case of research units, the size of the acquired data is usually quite limited, compared with datasets from private companies. On the downside, datasets from larger companies are typically unavailable for research purposes, as each owner is obliged to protect their users' data and releasing anonymized versions might even hurt the company's brand;
- **Combined approaches** are made up of the previous options, e.g. the initial simulation of data plus periodic updates with real measurements. This way, a large volume of data is ensured to create a repository and the data usable for processing will be gathered over time.

In addition to the above to ensure volume, other aspects are equally vital. **Data quality** control is crucial when gathering data from various crowdsourcing sources. The repository should operate only with trustworthy data relevant to the task at hand. Literature on data quality refers to relevant dimensions that should be considered when evaluating new data sources, such as accuracy, timeliness, consistency, quantity, relevancy, credibility, or interoperability [13] [14]. For the purpose of processing and analysis, data with adequate **time and spatial resolution** are required. This requirement needs to be evaluated in early phases of the project, as low data density leads often to inaccurate results. **Active and passive data collection** refers to the method of initiating the data measurement. In active mode the user (as in participatory sensing) initiates the action, while in passive mode it is initiated by software. Passive data collection seems a more appropriate method to avoid error-prone data and collect data in the correct resolution.

The **Data Storage** stage can be seen as a combination of hardware and software. **Hardware solutions** involve several approaches for the physical data storage such as storage server, internet cloud storage, or distributed storage system. The final choice depends on many factors, such as the availability of devices, financial resources, and even personal preferences. **Software solutions**, though, are specifically designed to cope with voluminous data. Examples are Apache Hadoop and Apache Spark, which are open-source big data processing platforms, supporting a variety of features such as distributed processing, libraries and third party modules for efficient data storage and management. **Data management** needs to be specially mentioned as a vital part of software solutions for data storage. It allows smooth and efficient handling and update of data in

repositories and databases. The existing solutions include MongoDB and Apache Atlas. Finally, **database size reduction** techniques allow to reduce the costs associated with storage requirements at the expense of reducing the quality of data or requiring extra effort when processing. That is, while big data file formats optimize storage and processing speed of large databases, data thinning decreases the frequency or resolution of data, making it more compact while reducing the quality to the predefined level. Data compression can lead to significant size reduction, yet reading the data requires decompression, slowing consequently overall performance. Modern compression methods can improve the signal to noise ratio of the data while significantly reducing the size of data [15, 16].

The **Data Processing** stage can involve visualization utilities, data mining algorithms, or data reduction methods. Methods for extrapolating new information from data can also be used as well as methods for data optimization for enhanced storage. An overview of available open source solutions for big data processing can be found in [17]. There are different strategies, namely:

- **Pre-processing** involves data wrangling operations to prepare data for further analysis, including format adjustment, compression, outliers detection and redundancy removal;
- **Analytics** software extracts quantitative and statistical information from the data. There are analytics oriented modules for Apache Spark [18], as well as stand-alone, compatible solutions such as Apache Zeppelin and Apache Kudu for Hadoop solutions;
- **Data mining** extracts hidden information by exploring the relations between variables. Data mining algorithms often utilize clustering, artificial neural networks and other machine learning methods. A remarkable open source Java-based solution is Rapid Miner, which is widely used in research, business and education. The Python-based Orange software contains lots of machine learning modules. Within the Apache family products, Apache Mahout and Apache SAMOA (for distributed machine learning applications) stand out;
- **Visualization** is an important step to better understand and reflect on the insights derived from a dataset. Data visualisation tools, which are often implemented in data processing frameworks, are then able to work in real-time with datasets and databases.

The challenges related to **Data Transmission** are also varied. **Connectivity technology** covers a wide variety of solutions with different data rates, power consumption, mesh types or range [19]. The optimal technology differs per device depending on the requirements and utilization. **Efficiency** of the data transfer, which is closely bound to the connectivity technology, depends on e.g. modulation, coding, and overheads due to data protection. **Encryption** requires greater processing power with higher levels of encryption. That's why small units with low computing and transmission power generally cannot afford high security during data transfer, although there are promising results in this area [20]. **Privacy** is still poorly addressed, with very little regulation from governments and regulatory bodies. As wearable data providers often collect significant amounts of personal data, filtering it would be unethical and cause global distrust toward the provider.

5 Conclusion

This paper surveys existing databases with crowdsourced data from wearables and discusses the main challenges related to data collection, storage, transmission and processing of wearable-based data. Based on this survey, we can state that efficiency, low computational power/low complexity and security are needed at all of these stages. Further research will focus on finding efficient/lossless compressing solutions of large datasets of wearable data to reduce delays in data transmissions, increase storage capacity, and accelerate data processing.

Acknowledgements

The authors gratefully acknowledge funding from European Union's Horizon 2020 Research and Innovation programme under the Marie Sklodowska Curie grant agreement No. 813278 (A-WEAR, http://www.a-wear.eu/).

References

[1] F. Delmastro, V. Arnaboldi, and M. Conti, "People-centric computing and communications in smart cities," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 122–128, 2016.

[2] R. Rawassizadeh, M. Tomitsch, M. Nourizadeh, E. Momeni, A. Peery, L. Ulanova, and M. Pazzani, "Energy-efficient integration of continuous context sensing and prediction into smartwatches," *Sensors*, vol. 15, no. 9, pp. 22616–22645, 2015.

[3] I. Torre, O. R. Sanchez, F. Koceva, and G. Adorni, "Supporting users to take informed decisions on privacy settings of personal devices," *Pers. and Ubiquitous Computing*, vol. 22, no. 2, pp. 345–364, 2018.

[4] W. Ugulino, D. Cardador, K. Vega, E. Velloso, R. Milidiú, and H. Fuks, "Wearable computing: Accelerometers' data classification of body postures and movements," in *Brazilian Symposium on Artificial Intelligence*, pp. 52–61, Springer, 2012.

[5] W.-H. Lee, J. Ortiz, B. Ko, and R. Lee, "Time series segmentation through automatic feature learning," *arXiv preprint arXiv:1801.05394*, 2018.

[6] M. Aazam, M. St-Hilaire, C. Lung, and I. Lambadaris, "Pre-fog: Iot trace based probabilistic resource estimation at fog," in *2016 13th IEEE Annual Consumer Communications Networking Conference (CCNC)*, pp. 12–17, Jan 2016.

[7] D. Jarchi and A. Casson, "Description of a database containing wrist ppg signals recorded during physical exercise with both accelerometer and gyroscope measures of motion," *Data*, vol. 2, no. 1, p. 1, 2017.

[8] J. Wang, Y. Chen, L. Hu, X. Peng, and S. Y. Philip, "Stratified transfer learning for cross-domain activity recognition," in *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–10, IEEE, 2018.

[9] D. Morris, T. S. Saponas, A. Guillory, and I. Kelner, "Recofit: using a wearable sensor to find, recognize, and count repetitive exercises," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3225–3234, ACM, 2014.

[10] "Github open-access data repository." https://github.com.

[11] S. Liu, S. Schiavon, H. P. Das, M. Jin, and C. J. Spanos, "Personal thermal comfort models with wearable sensors," *Building and Environment*, vol. 162, p. 106281, 2019.

[12] "Data.gov on-line repository." https://data.gov/.

[13] I. Taleb, M. A. Serhani, and R. Dssouli, "Big data quality: A survey," in *2018 IEEE International Congress on Big Data (BigData Congress)*, pp. 166–173, IEEE, 2018.

[14] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data science journal*, vol. 14, 2015.

[15] J. Talvitie, M. Renfors, and E. S. Lohan, "Novel indoor positioning mechanism via spectral compression," *IEEE Communications Letters*, vol. 20, no. 2, pp. 352–355, 2015.

[16] J. Talvitie, M. Renfors, M. Valkama, and E. S. Lohan, "Method and analysis of spectrally compressed radio images for mobile-centric indoor localization," *IEEE Transactions on Mobile Computing*, vol. 17, no. 4, pp. 845–858, 2017.

[17] A. Londhe and P. P. Rao, "Platforms for big data analytics: Trend towards hybrid era," in *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, pp. 3235–3238, IEEE, 2017.

[18] G. Gousios, "Big data software analytics with apache spark," in *2018 IEEE/ACM 40th International Conference on Software Engineering: Companion (ICSE-Companion)*, pp. 542–543, IEEE, 2018.

[19] T. Liang and Y. J. Yuan, "Wearable medical monitoring systems based on wireless networks: A review," *IEEE Sensors Journal*, vol. 16, no. 23, pp. 8186–8199, 2016.

[20] H. Tahir, R. Tahir, and K. McDonald-Maier, "On the security of consumer wearable devices in the internet of things," *PloS one*, vol. 13, no. 4, p. e0195487, 2018.