Università di Genova | DIPARTIMENTO INFORMATICA, BIOINGEGNERIA, ROBOTICA E INGEGNERIA DEI SISTEMI

Computer Science Workshop
PhD program in Computer Science and Systems Engineering

Psst…Hey! Do you have some time to talk about reproducibility of your work?

Andrea Fasciglione, Maurizio Leotta, Alessandro Verri

Reproducibility in Activity Recognition based on Wearable Devices: a Focus on used Datasets

Motivation & Context

Reproducibility in proposed approaches is a crucial element in scientific fields. Concentrating on the task of human activity recognition (AR) using data coming from wearable devices with inertial sensors, we have analyzed the reproducibility of proposed approaches with a focus on used datasets. We have measured what percentages of works verified their approach using public datasets or sharing the ones created on purpose. Moreover, we have examined the characteristics of considered datasets, with attention to involved populations and recorded activities.

Our goal is to bring the attention of other researchers to these issues and to persuade future articles' authors to support the replicability and reproducibility of their works.

Analyzed Aspects

Starting from 1289 articles retrieved on Scopus Digital Library, we have analyzed in details 146 of them. We selected these remaining works by considering only works that had a number of citations per year higher than 5.0 (gathering, in this way, 207 works). We then excluded other works using defined inclusion and exclusion criteria, in order to omit works not related to our interests (obtaining, in this way, 146 works).

We read and examined remaining candidate articles by filling out a detailed form consisting of different questions, elaborated in order to extract statistics about public access to datasets and their characteristics.

Hereby some of the questions of the form:

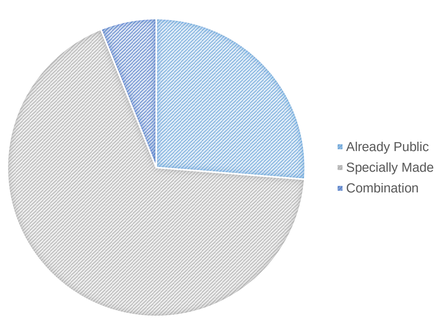
1. Was link (or access) to the used dataset provided at the time of paper publication?
2. Is the used dataset still available?
3. How many activities have been studied?
4. How many subjects have been involved in data recording?
5. What percentages of male & female participants have been involved?
6. What is the average age of considered subjects?
7. What are the average weight and height (or BMI) of subjects?

Specially Made vs Already Public Datasets

It is useful to firstly notice that researchers proposing a novel AR method can evaluate its effectiveness by using:
- one (or more) already public datasets
- one (or more) datasets specially made for that article
- a combination of already public and specially-made datasets

Out of 146 considered articles, we have found out that:
- 38/146 (~26%) were using already publicly available datasets
- 98/146 (~67%) were using datasets specially made for associated works
- 10/146 (~7%) were using a combination of already publicly available datasets and specially-made datasets

As we can see the majority of works (> 67%) evaluated the proposed approach with one or more datasets that have been specially recorded for that work. At the same time, only about a third of the works used already publicly available datasets.



Public Access to Datasets

Starting from these 146 articles, we have found 110 already publicly available datasets and 31 specially made datasets, for a total of 141 datasets

What about the accessibility of these 141 datasets?

	Free Access	Restricted Access*	Offline **	Not Shared
Already Public	27 / 31 (87%)	2 / 31 (6%)	2 / 31 (6%)	0 / 31 (0%)
Specially Made	16 / 110 (15%)	8 / 110 (7%)	7 / 110 (6%)	79 / 110 (72%)
Total	43 / 141 (30%)	10 / 141 (7%)	9 / 141 (6%)	79 / 141 (56%)

* In these cases, we could potentially access data after (a) contacting authors or (b) registering to specific websites (e.g. IEEE DataPort, requiring a subscription for a fee)
** In these cases, associated links or servers are nowadays not accessible; even a research using common web search engines could not help us to retrieve these datasets

Characteristics of Datasets

A first interesting aspect that we would like to point out is related to gathering information about considered populations used to record data: quite often this information was not given. In fact, for 31 datasets out of 141 (~22%), we could find none of the basic information of about gender distribution, ages, weights, and heights or BMIs of people involved in data recording.

In the following table the availability of information regarding datasets.

# Subjects	Gender	Age	Weight	Height	BMI	# Activities
140 / 141 (99%)	93 / 141 (66%)	102 / 141 (72%)	48 / 141 (34%)	45 / 141 (32%)	47 / 141 (33%)	140 / 141 (99%)

In the following table, the median values for some of the considered features of datasets and populations.

# Subjects	Gender	Age	Weight	Height	BMI	# Activities
10	61% (M) 39% (F)	29	70 Kg	171 cm	24 kg/m^2	8

The median age of subjects engaged is around 29 years, meaning that the considered populations are usually young, even if typical applications of AR systems concern elderly people.

Possible Advices

Starting from our findings, here we suggest possible advices for future papers' authors that desire to record and publish new datasets or that wish to propose novel approaches with specially made datasets.

1. Share used datasets and all the additional information that can be useful for readers to compare works and approaches
2. Adhere to Open Access principles when publishing articles, especially if presenting new datasets so that connected material can be openly accessible for any possible user
3. Take care of the population involved in data recording in order to combine subjects with heterogeneous characteristics

Forthcoming Research

1. Further analyze the reproducibility of approaches by analyzing more aspects of methods that should be shared (e.g. implementation details)
2. Check if there is any trend (positive or negative) in lack of reproducibility details over time, from the past up to now, for example by replicating our analyses clustering the works per different years.

Acknowledgements
This work is an activity of the Software Engineering for Healthcare Lab, a joint laboratory between the University of Genova and Janssen Pharmaceuticals (Johnson & Johnson group). We want to show our gratitude to Massimo Raineri, Giacomo Ricca, and Salvatore Saporito of Janssen Pharmaceuticals for the support provided.

REFERENCE ARTICLE

Andrea Fasciglione, Maurizio Leotta and Alessandro Verri, «*Reproducibility in Activity Recognition Based on Wearable Devices: a Focus on Used Datasets*», currently under review for IEEE International Conference on Systems, Man and Cybernetics.

CONTACTS

Andrea Fasciglione
andrea.fasciglione@edu.unige.it

Maurizio Leotta
maurizio.leotta@unige.it

Alessandro Verri
alessandro.verri@unige.it

OUR PUBLIC DATASET

Do you want to use our dataset for activity recognition? ;)

You can find it at the following links:

- https://sepl.dibris.unige.it/2020-DailyActivityDataset.php
- https://doi.org/10.7910/DVN/G23QTS

…. Or by following the QR code! ☺

CSW