Potential Thesis

Statistically, Appropriate Method To Evaluate The Impact Of The Subject Selection Bias On HAR

Datasets are created to help researchers analyze, develop and evaluate methods purposed for solving real problems. However, dataset creation is a tedious process. Aspects such as sample size, control environment, sensors, etc., need to be evaluated and documented to ensure that the created dataset will be informative and expandable for future research. Consequently, dataset creation is of interest for the research community associated with Human Activity Recognition (HAR). The past has seen the HAR community invest in the creation of benchmarked, open datasets for method evaluation.

HAR is the task of classifying human movements based on the activity performed [1]. Typically, signals from videos or multi-channel time-series, eg: measurements from on-body devices such as inertial measurement units (IMUs), are used to perform classification. Over the years, various types of classifiers, kNN, SVM, Decision tree, Deep Neural Networks (DNNs), etc., have been used to perform classification. HAR is of interest because of the non-determinist nature of the human movement. For instance, when an individual repeats a task, the sensor patterns are not identical. This implies that HAR faces repetition without repetition. In [1], a logistics activity-based dataset was created for HAR as part of the DFG Project, Transfer Learning for Human Activity Recognition in Logistics (Fi799/10-2 & HO2403/14-2). On experimenting on the dataset to analyze whether person re-identification can be performed irrespective of the activity, optimistic results were gathered, [2]. Furthermore, it was identified that human characteristics such as age, gender identity, height, etc., can be identified from the HAR datasets. This implies that each individual has a unique motion behavior when performing activities, influenced by their physical characteristics. The classifier used for the experiments was a DNN – CNN-IMU network [3].

Often, HAR datasets are created with subjects who are immediately available to the dataset creators, for example, students or colleagues. The impact of the selected subjects performing the activity is rarely discussed. Given that, the subject's individuality is present within the HAR dataset, it is expected that the individual's idiosyncrasy has an impact on HAR. This understanding, thus, brings us to the questions:

1. How do the dataset's subject characteristics affect HAR?
2. What is the appropriate method to evaluate the effect?

In this collaborative project between the Department of Statistics and Chair of Material Handling and Warehouse at TU Dortmund, we attempt to bring out the statistically appropriate method to evaluate the impact of the subject selection bias on HAR.

The HAR dataset is not devoid of biases. From sensor position bias, to feature selection bias, researchers have often focused on accomodating the biases in the classifier model, [4], [5], [6]. Others have attempted at modifying the train-validation method to overcome the impact of bias, [7]. However, as part of this thesis, we are interested to approach the problem of bias from the perspective of the dataset. Consequently, we attempt to find a statistically appropriate method to evaluate which are the physical characteristics that affect/bias HAR. The study is expected to help define a rule of thumb in subject selection for the creation of HAR datasets, similar to [8].

Reference

[1] F. Niemann *et al.*, “LARa: Creating a Dataset for Human Activity Recognition in Logistics Using Semantic Attributes,” *Sensors*, vol. 20, no. 15, p. 4083, Jul. 2020, doi: 10.3390/s20154083.

[2] N. R. Nair, Person Identification Using Motion Information, Master Thesis, 2021. URL: https://patrec.cs.tu-dortmund.de/pubs/theses/ma_nair.pdf

[3] Moya Rueda, Fernando, et al. "Convolutional neural networks for human activity recognition using body-worn sensors." *Informatics*. Vol. 5. No. 2. Multidisciplinary Digital Publishing Institute, 2018.

[4] Tzeng, Eric, et al. "Deep domain confusion: Maximizing for domain invariance." *arXiv preprint arXiv:1412.3474* (2014).

[5] Hamidi, Massinissa, and Aomar Osmani. "Description of Structural Biases and Associated Data in Sensor-Rich Environments." *arXiv preprint arXiv:2104.04885* (2021).

[6] Hamidi, Massinissa, and Aomar Osmani. "Human Activity Recognition: A Dynamic Inductive Bias Selection Perspective." *Sensors* 21.21 (2021): 7278.

[7] Koskimäki, Heli. "Avoiding bias in classification accuracy-a case study for activity recognition." *2015 IEEE symposium series on computational intelligence*. IEEE, 2015.

[8] Riley, Richard D., et al. "Calculating the sample size required for developing a clinical prediction model." *Bmj* 368 (2020).