Datasheet for PsyMo: A Dataset for Estimating Self-Reported Psychological Traits from Gait

Adrian Cosma
University Polidrtehnica of Bucharest
Bucharest, Romania
cosma.i.adrian@gmail.com

Emilian Radoi
University Politehnica of Bucharest
Bucharest, Romania
emilian.radoi@upb.ro

1. Motivation For Datasheet Creation

1.1. Why was the datasheet created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

This datasheet was created to accompany the WACV 2024 submission for the PsyMo dataset, and details the composition, collection process, distribution, preprocessing, maintanance and ethical considerations for the PsyMo dataset. PsyMo is intended for exploration of psychological manifestations in walking patterns. PsyMo can be used for benchmarking models in the estimation of 17 psychometric attributes from gait in multiple variations.

1.2. Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?

The dataset has not been currently used in other works. We invite researchers to propose baseline results for the psychological trait estimation and gait recognition tasks mentioned in the article.

1.3. What (other) tasks could the dataset be used for?

PsyMo's main purpose is estimation of psychometric attributes from gait. However, due to its size and controlled diversity of walking variations and viewpoints, it could be used to benchmark models in gait recognition.

1.4. Who funded the creation dataset?

The dataset creation received no external funding. All subjects in the dataset were volunteers.

2. Datasheet Composition

2.1. What are the instances?(that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

PsyMo contains 312 different subjects walking captured in different variations and camera viewpoints. We process the walking sequences using a state-of-the-art AlphaPose [1] to obtain skeleton sequences, CLIFF [?] to estimate 3D human pose and 3D meshes in the form of parametric SMPL predictions, and extracted silhouettes using pretrained instance segmentation model (Hybrid Task Cascade [2]) to obtain silhouette sequences. Skeletons are composed of 18 joint coordinates in the image plane, with x and y coordinates and an additional confidence score for each joint, which measures detection quality. Each sequence is provided in JSON format. SMPL information is provided in .npy files, with the same format as in the opensource implementation [1]. Additionally, each silhouette is provided in a 128x128 image in .PNG format. Each silhouette is centered in the frame. We also provide Gait Energy Images (GEI), for convenience, in .PNG format. GEI's are constructed by averaging the silhouettes of a walk.

2.2. How many instances are there in total (of each type, if appropriate)?

The are a total of 14,976 walking sequences, from 312 individuals. Each subject has 48 walks, across 6 viewpoints (0°, 45°, 90°, 180°, 225°, 270° including round-trips) and 7 walking variations (normal walking, carrying bag, clothing change, walk speed slow, walk speed fast, talking on the phone and texting while walking). For normal walking, subjects walked a total of 4 times, while on other variations each subject walked 2 times.

[1] https://github.com/huawei-noah/noah-research/tree/master/CLIFF

2.3. What data does each instance consist of ? \Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are sub-populations identified (e.g., by age, gender, etc.) and what is their distribution?

Skeleton sequences and SMPL parameters are not further processed, and are provided as obtained by AlphaPose and CLIFF, respectively. For silhouettes, raw pixel probabilities are provided, but the silhouettes are centered in the image and rescaled to 128x128. Each subject is associated with age, gender, weight and height information, alongside raw scores and ordinal classes for 17 psychological attributes.

2.4. Is there a label or target associated with each instance? If so, please provide a description.

Each walking sequence has information regarding the viewpoint and walking variation. For each subject, there is information regarding their age, gender, weight, height. Additionally, raw scores 6 psychological questionnaires are provided, totalling 17 psychological attributes, across factors and subscales. The questionnaires are the Big Five Index (BFI) [3], Rosenberg Self-Esteem (RSE) [4], Buss-Perry Aggression Questionnaire (BPAQ) [5], Ocupational Fatigue Exhaustion / Recovery Scale (OFER) [6], Depression, Anxiety and Stress Scale (DASS) [7] and General Health Questionnaire (GHQ) [8]. We also provide ordinal classes for each subscale / factor. For BFI and BPAQ we obtain ordinal classes using quantiles on the raw score, while on the others we used their respective threshold values.

2.5. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

There is no missing information for each instance / subject.

2.6. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

There are no relationships between subjects or walking sequences. Each walking sequence has explicit information regarding viewpoint and walking variation.

2.7. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

PsyMo does not cover all possible viewpoints and walking variations, and is instead collected in a controlled environment to benchmark performance on specific variations, available for all subjects in the dataset, similar to other datasets in this domain [9–11]. Datasets such as DenseGait [12] and GREW [13] provide ample diversity of walking variations present in the real world, but fine-grained annotation with psychological attributes is unfeasible. All subjects contained in PsyMo are Romanian, however it is a representative sample for a proof of concept on estimating psychometric attributes from walking.

2.8. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

For all tasks, we recommend a 80:20 training / validation split on the subjects, corresponding to 250 subject for training and 62 subject for validation.

2.9. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

The skeleton provided by the pose estimation model might not always be correctly extracted, due to occlusion of joints in some viewpoints. However, we did not address this issue (for example, by using dedicated hardware for skeleton estimation such as Kinect) as this would not be available in realistic surveillance scenarios. As such, approaches that process PsyMo for a particular task may also include a way to rectify skeletons. The same is true for the extracted silhouettes and SMPLs.

2.10. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained.

3. Collection Process

3.1. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

We used three Tapo C200 [2] consumer surveillance cameras, for their ease of use, widespread availability and web API [3]. The cameras was synchronized and controlled using a custom built python program. Subjects filled in 6 psychological questionnaires remotely using Google Forms, alongside attributes related to their body composition (age, gender, weight, height). Questionnaires are validated on large-scale populations, and have official translations in Romanian; if there are no translations, we translated them with the help of a psychologist.

3.2. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data associated with each instance is self-reported individually by each subject through survey responses. Questionnaires have been validated in literature and contain redundancies both in terms of individual items in each questionnaire. Moreover, we have some redundancies / correlations across questionnaires, see Figure 1)

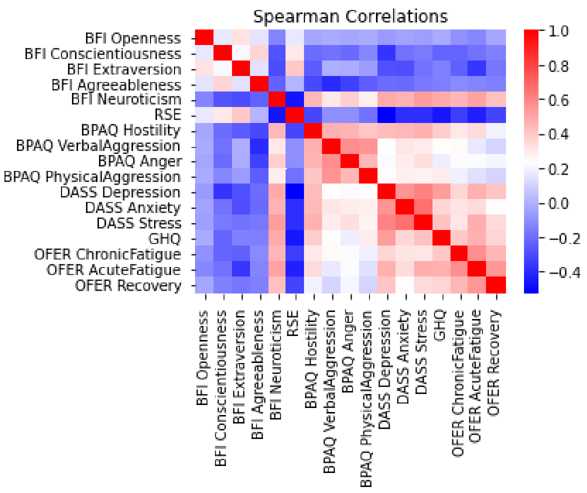

Figure 1. Spearman correlations between questionnaire scores.

3.3. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The dataset is not sampled from a larger set.

3.4. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The dataset is comprised of student volunteers.

3.5. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The dataset was collected over the course of 2 months. Each subject was required to fill in the psychological questionnaires and then to present themselves to the lab to have their walk captured.

4. Data Preprocessing

[2] https://www.tp-link.com/ro/home-networking/cloud-camera/tapo-c200/

[3] https://github.com/JurajNyiri/pytapo

4.1. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

There has not been any significant preprocessing done on the dataset, except for the calculation of the final scores on each questionnaires, according to the respective specification for each questionnaire.

4.2. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

N/A

4.3. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

N/A

4.4. Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?

N/A

4.5. Any other comments

5. Dataset Distribution

5.1. How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

Currently, the dataset is distributed at https://bit.ly/3Q91ypD.

5.2. When will the dataset be released/first distributed? What license (if any) is it distributed under?

The dataset is available immediately at https://bit.ly/3Q91ypD. The dataset is protected by the CC-BY-NC-ND[4] License.

5.3. Are there any copyrights on the data?

The dataset does not contain any copyrighted content, and was collected entirely by the authors. The dataset is protected by copyright through the CC-BY-NC-ND License.

5.4. Are there any fees or access/export restrictions?

There are no fees or restrictions.

6. Dataset Maintenance

6.1. Who is supporting/hosting/maintaining the dataset?

The dataset is supported, hosted and maintained by the authors.

6.2. Will the dataset be updated? If so, how often and by whom?

If there is a rationale for updating the dataset (e.g. extending or correcting it), the authors will make the necessary modifications.

6.3. How will updates be communicated? (e.g., mailing list, GitHub)

Updates will be posed on the hosting website.

6.4. If the dataset becomes obsolete how will this be communicated?

In case the dataset will become obsolete, it will be rectracted and an update will be posted on the hosting website.

6.5. Is there a repository to link any/all papers/systems that use this dataset?

We will make a dedicated webpage on the hosting website which will feature any system that uses PsyMo.

6.6. If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

N/A. Currently PsyMo is not intended to be extended by third parties except the authors.

7. Legal and Ethical Considerations

[4] https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode

7.1. Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

The collection of PsyMo has been approved by the Ethics Review Board. The approval documentation and a complete description of the data collection procedure will be made available after the anonymization period.

7.2. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No, the dataset does not contain confidential information. All data is self-reported by each participant under explicit and informed consent.

7.3. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why

No, PsyMo does not contain offensive information, it is comprised of walking sequences (skeletons and silhouettes), annotated with 17 psychometric attributes.

7.4. Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset is comprised of walking sequences from 312 volunteer subjects.

7.5. Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these sub-populations are identified and provide a description of their respective distributions within the dataset.

Of the 312 participants, 113 were female and 199 were male, with an average age of 21.9 years (SD = 2.18). Moreover, the average weight for the participants is 70.5kg (SD = 15.7) with the average height of 174.8cm (SD = 8.9), corresponding to an average BMI of 22.87 (SD = 3.9).

7.6. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

The dataset is anonymized: we do not release the raw walking videos, only anonymized skeleton sequences and silhouette sequences. It is not possible to identify subjects unless their walking sequence is annotated with their identity.

7.7. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

The dataset does not contain any data related to the ones enumerated above. PsyMo contains responses from 6 psychological questionnaires, two of which related to mental health (DASS-21 and GHQ-9). However, they were provided under explicit and informed consent, and any identifiable information has been removed; PsyMo is anonymized.

7.8. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

All data was directly self-reported by each participant. Each walking instance is directly captured with explicit and informed consent in predetermined variations.

7.9. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

We announced our intention of collecting this dataset with the following message:

"This assignment represents an opportunity for you to contribute to the research performed in our Computer Science department of our university, by helping us collect a dataset which will enable an interdisciplinary study on personality traits and movement. This requires you filling in 6 personality questionnaires and walking multiple times in front of three cameras."

Students voluntarily participated in this study in their own terms, with full knowledge of the dataset collection procedure, distribution and intended purposes. The dataset collection was approved by the Ethical Review Board.

7.10. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Before filling in the questionanires and having their walk captured, subjects were prompted for their consent, after a description of the study (translated to English from Romanian):

This research is carried out within the Department of Computers Science. Your participation will help us explore the possible correlations between psychometric information and human movement in physical space. This is a pilot study of an unexplored area in the field of psychology and artificial intelligence. The aim of this study is to investigate the manifestations of psychometric attributes on gait in different scenarios that simulate real gait situations (normal gait, different clothing, backpack etc).The study consists of completing a set of psychometric questionnaires that measure a set of attributes related to personality and psychological disorders. After completing the questionnaires, your walking patterns will be recorded in different scenarios after an appointment.

In accordance with the requirements of Regulation (EU) 2016/679 on the protection of individuals with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation) and Law no. 506/2004 on the processing of personal data and the protection of privacy, the research team has the obligation to manage safely and only for the specified purposes. The data you will provide: demographic data, answers to questionnaires and movement information. The statistical processing of the provided data will be analyzed at the sample level and will not be presented individually in any scientific publication. The recorded information will only be used by members of the research team. After the collection process is completed, the data will be anonymized, of interest being only the movement information. The results of the research will be made public only in an anonymized version, without being able to reach the identity of the people present in the study. Your participation in this research is entirely voluntary. By choosing to participate you agree to the processing of personal data provided for research purposes. If you have any questions, concerns, or complaints, or if you would like to report any research-related harm or abuse, please contact us.

7.11. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Subjects can revoke their consent by directly contacting the authors via email.

7.12. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

A data protection impact analysis has not been performed on PsyMo. The dataset does not contain sensitive information and is anonymized.

References

[1] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*, 2018. 1

[2] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019. 1

[3] Oliver P John, Sanjay Srivastava, et al. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138, 1999. 2

[4] Morris Rosenberg. *Society and the Adolescent Self-Image*. Princeton University Press, 2015. 2

[5] Arnold H Buss and Mark Perry. The aggression questionnaire. *Journal of personality and social psychology*, 63(3):452, 1992. 2

[6] PC Winwood, AH Winefield, D Dawson, and K Lushington. Development and validation of a scale to measure work-related fatigue and recovery: the occupational fatigue exhaustion/recovery scale (ofer). *Journal of Occupational and Environmental Medicine*, pages 594–606, 2005. 2

[7] Sydney H Lovibond and Peter F Lovibond. *Manual for the Depression Anxiety Stress Scales, second ed.* Sydney: Psychology Foundation, 1995. 2

[8] David P Goldberg and Barry Blackwell. Psychiatric illness in general practice: a detailed study using a new method of case identification. *Br med J*, 2(5707):439–443, 1970. 2

[9] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 441–444. IEEE, 2006. 2

[10] Ziyuan Zhang, Luan Tran, Feng Liu, and Xiaoming Liu. On learning disentangled representations for gait recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, Sep. 2019*, June 2019. 2

[11] Y. Makihara, H. Mannami, A. Tsuji, M.A. Hossain, K. Sugiura, A. Mori, and Y. Yagi. The ou-isir gait database comprising the treadmill dataset. *IPSJ Trans. on Computer Vision and Applications*, 4:53–62, Apr. 2012. 2

[12] Adrian Cosma and Emilian Radoi. Learning gait representations with noisy multi-task learning. *Sensors*, 22(18), 2022. 2

[13] Zheng Zhu, Xianda Guo, Tian Yang, Junjie Huang, Jiankang Deng, Guan Huang, Dalong Du, Jiwen Lu, and Jie Zhou. Gait recognition in the wild: A benchmark. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2