

Datasheet Version 1.0

Title	Details
1. Reason for Creation	This dataset was created to enable the training of AI systems that can recognize identifiable professionals.
2. Existing use of the dataset	The dataset has not been deployed for practical use as at the time of preparing this datasheet.
3. Source of Funding for the creation of the dataset	The collection and preparation of this dataset as been purely funded by personal funds of the collector.
4. Examples contained in the dataset	The dataset contains images of individuals dressed in their respective professional uniforms.
5. Number of images in the dataset	There are 11,000 images contained in the dataset that covers 10 professional examples. Each example has 1,100 images, that is divided into 900 training and 200 test images.
6. Example composition	Each professional example in the dataset is contained in a separate folder, with the folder name corresponding to the example label (e.g doctor). Each image in each example is a 224 * 224 pixel JPEG image.
7. Experiments conducted on the dataset	A popular machine learning network has been trained on the dataset and the results are available in the GitHub repository of the dataset. https://github.com/OlafenwaMoses/IdenProf
8. Year of release	2018
9. Number of images contained	11,000
10. Size of each image	224 * 224 pixels
11. Percentage of female subjects	19.4 %
12. Percentage of male subjects	80.6 %
13. Percentage of white subjects	91.1 %
14. Percentage of black subjects	8.9 %
15. Process of dataset collection	The images in the dataset were obtained from Google Image search. The images were searched and collected based on the 15 most populated countries in the world.
16. Team/Personals that collected the dataset	The dataset was compiled by Moses Olafenwa.
17. Period of data collection	The dataset was collected over a period of 3 months.
18. Relation of dataset with image instances	The personals whose images are contained in the dataset are identified by their profession.
19. If dataset contain all possible categories	The dataset does not contain all possible categories of professions. Work is ongoing to expand the dataset and add more categories.

Title	Details
20. Processing of the image samples	<p>The images contained in the dataset set were:</p> <ul style="list-style-type: none"> • Obtained from Google Images search • Sorted with false positives and repetitions removed • Arranged to respective category folder • Resized to 224 * 224 pixel
21. Preservation of raw/unprocessed images	The unprocessed images are preserved.
22. Challenges collecting the dataset	<p>In collecting the dataset, the following challenges were encountered:</p> <ul style="list-style-type: none"> • Obtaining equal number of image instances per category kept the number of that can be supported now at 10 • There were more images of Asians and Whites obtainable than Blacks • There were more images of men obtainable than women
23. Dataset distribution	The dataset is available in a zip format.
24. Release date	The dataset was released in May 2018.
25. License	<p>The dataset is released under the MIT License . The details are available in the GitHub repository.</p> <p>https://github.com/OlafenwaMoses/IdenProf</p>
26. Fees of Restriction	There are no fees or restrictions in the use of the dataset.
27. Project Support	Project support is provided by the “AI Commons” teams, that includes Moses and John Olafenwa.
28. Project Hosting	<p>The project will be hosted and available via the IdenProf GitHub repository.</p> <p>https://github.com/OlafenwaMoses/IdenProf</p>
29. Update and new Releases	<p>All updates will be made available via the IdenProf GitHub repository.</p> <p>https://github.com/OlafenwaMoses/IdenProf</p> <p>Work is on to ensure IdenProf is updated for corrections every 6 months and new versions are provided twice a year.</p>
30. Informing and Obtaining consent from individuals contained in the sample images	<p>The individuals whose images are contained in the images in this dataset were not contacted in regards to the use of their images for the dataset. The images were however obtained from the open web. Also, we can’t guarantee all the images are free from copyrights. We recommend that complaints of copyright observed in the dataset be submitted as an issue to the GitHub repository.</p> <p>https://github.com/OlafenwaMoses/IdenProf</p>

Title	Details
31. Exposing individuals to risk	The dataset in no way exposed the individuals whose images are contained to any risks, as the images were obtained from the publicly accessible web.
32. Compliance with EU General Data Protection Regulation (GDPR)	The dataset does not comply with EU GDPR as the individuals whose images were contained were not explicitly contacted for consent.
33. Containing confidential details	The dataset does not contain any form of confidential details on any of the individuals whose images are contained.

Prepared by : Moses Olafenwa
Email : guymodscientist@gmail.com
GitHub: https://github.com/OlafenwaMoses
Date : 18th of July, 2018