Datasheet for the Faces Through Time Dataset

Motivation

Why was the dataset created?

Faces Through Time was created to facilitate future academic Computer Vision research involving the appearances of people across the 19th, 20th and 21st centuries.

Who created this dataset (*e.g.*, which team, research group) and on behalf of which entity (*e.g.*, company, institution, organization)?

The dataset was created by researchers at Cornell University.

Who funded the creation of the dataset?

The creation of the dataset was funded by Cornell University. Please see the Acknowledgements section for more information.

Composition

What do the instances that comprise the dataset represent (*e.g.*, documents, photos, people, countries)? Are there multiple types of instances? (*e.g.*, movies, users, ratings; people, interactions between them; nodes, edges)

The instances are cropped facial images.

Are relationships between instances made explicit in the data (*e.g.*, social network links, user/movie ratings, etc.)?

No.

How many instances are there? (of each type, if appropriate)?

26,247 images (corresponding to 26,247 unique individuals).

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 subpopulations identified (*e.g.*, by age, gender, etc.) and what is their distribution?

Each instance consists of an aligned and resized facial image (256×256 pixels).

We can extract biographic labels per identity from Wikimedia Commons, such as birth year and gender. We only release the year a photo was taken in and these two biographic labels with each (anonymous) facial image.

The other available biographic labels include citizenship, occupation, spoken language, and ethnic groups (some subset of these labels may be available for each identity). To illustrate the demographics in our dataset, we report the number of identities with labels under each biographic attribute in Table 1. In Figure 1, we plot a histogram of the birth years in our dataset. However, we do not release these additional labels with our data.

Attribute	Identities	Labels
Occupations	16,692	1,031
Citizenships	16,411	259
Languages	7,564	117
Ethnic Groups	956	125

Table 1: Number of identities with labels under each biographic attribute and the number of unique labels for each attribute.

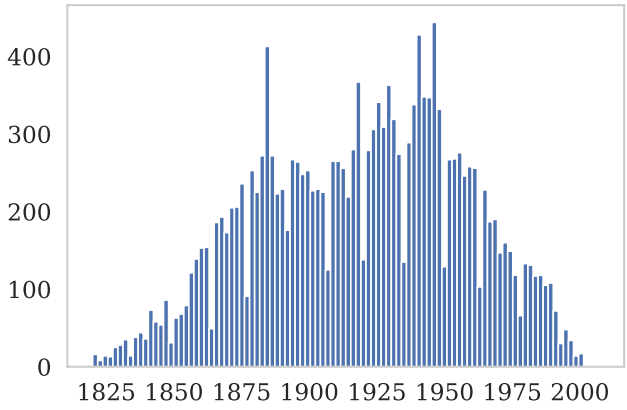

Figure 1: Histogram of birth years in *Faces Through Time*.

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.

Yes, as detailed in Table 1, individuals may only contain a subset of the mentioned biographic labels, according to their availability on Wikimedia Commons.

In addition, we anonymize our dataset by removing each identity's associated name and original Wikimedia Commons source (and biographic information is not released with the dataset).

Is everything included or does the data rely on external resources? (*e.g.*, websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version; c) are there access restrictions or fees?

The dataset is self-contained.

Are there recommended data splits and evaluation measures? (*e.g.*, training, development, testing; accuracy or AUC)

We randomly select 100 images from each decade, totaling 1,400 images, for the *Faces Through Time* test set. The remaining images are used for training.

Are there any errors, sources of noise, or redundancies in the dataset?

Timestamps provided for each image are somewhat noisy. As detailed in the paper, we compare it against the individual's birth year.The images associated with an individual were selected using a automatic clustering technique (as described in the paper). Our method is highly accurate (yielding 93% accuracy on a 50 identity test set), however, as with most automatic algorithms, it is not completely error-free.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (*e.g.*, websites, tweets, other datasets)?

The dataset is self-contained.

Does the dataset contain data that might be considered confidential (*e.g.*, data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)?

No. All data was derived from Wikimedia Commons. Images are freely-licensed and provided through a public catalog.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

As we describe in our paper, we don't believe so, due to the nature of our data source (Wikimedia Foundation projects). However, this will require much closer auditing, so we currently have insufficient information to determine this.

Does the dataset relate to people?

Yes. This is a people-centric dataset that contains facial images.

Does the dataset identify any subpopulations (*e.g.*, by age, gender)?

The information is available on Wikimedia Commons, which identifies subpopulations by age, gender, citizenship, spoken language, occupation, and ethnic group. The distributions for occupations, citizenships, and languages are reported in Figure 2, Figure 3, and Figure 4 respectively. However, this information will not be available in the dataset, and the image samples are anonymous.

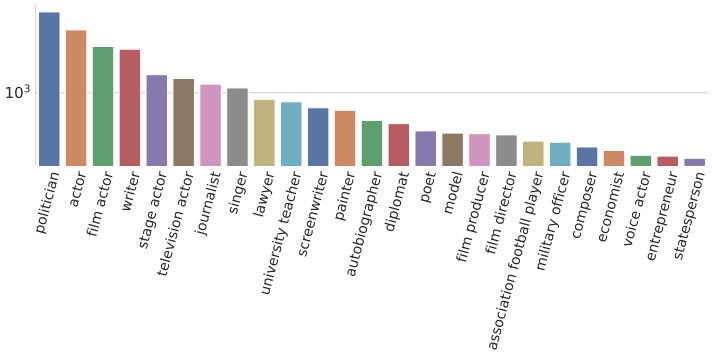

Figure 2: Distribution of top 25 most common occupations in *Faces Through Time*.

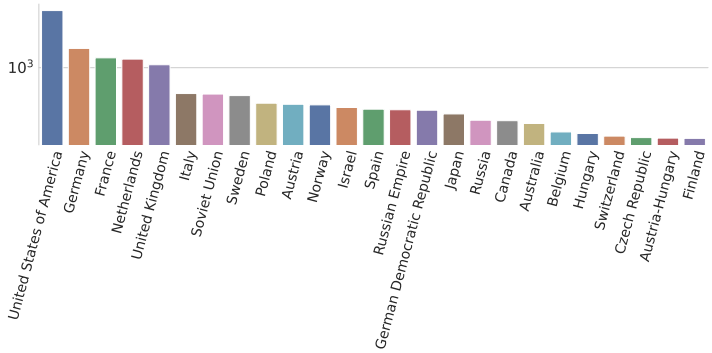

Figure 3: Distribution of top 25 most common citizenships in *Faces Through Time*.

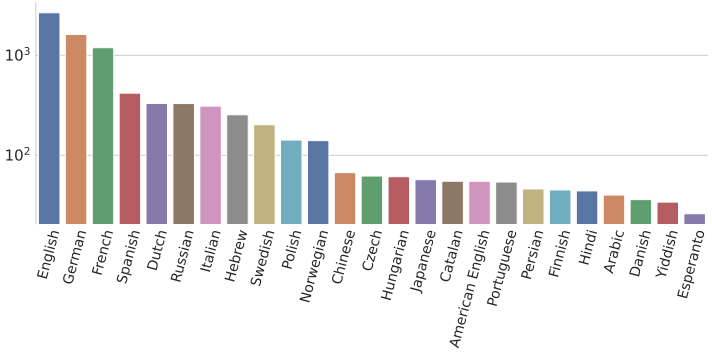

Figure 4: Distribution of top 25 most common languages spoken in *Faces Through Time*. Note that only a third of identities are tagged with languages spoken.

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?

It is not possible to identify individuals from the our dataset.

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)?

The information may be available on Wikimedia Commons. However, this information will not be available in the dataset, and the image samples are anonymous.

Any other comments?

Data Collection Process

How was the data associated with each instance acquired?

The data was acquired under the "People by name" category in Wikimedia Commons, which contains an instance name and multiple images associated with that individual. As described in the text, we cluster each collection to obtain a clean subset which captures this individual. We scrape additional samples without biographic labels from the "19th Century Portrait Photographs of Women" and "20th Century Portrait Photographs of Women" categories. These make up about 15% of the dataset. We then group images by decades, according to the images' timestamps.

What mechanisms or procedures were used to collect the data (*e.g.*, hardware apparatus or sensor, manual human curating, software program, software API)?

Automatic scraping procedures were used to collect the data.

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

Our clustering technique provided a collection of multiple images per identity. However, we intentionally limit the number of images to one per identity in the release version to prevent applications of facial recognition. The sampling strategy (selecting one image per identity) is probabilistic.

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

The authors of this paper were solely involved in the data collection process.

Over what time-frame was the data collected?

Our dataset reflects a recent state of Wikimedia Commons and Wikidata. Exact time-frame will be released together with the dataset.

Were any ethical review processes conducted (*e.g.*, by an institutional review board)?

We internally conducted ethical review (no official processes were conducted, due to the public nature of the data on Wikimedia Commons).

Does the dataset relate to people?

Yes. Each instance in the dataset represents a person.

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

The data was gathered from Wikimedia Commons and Wikidata.

Were the individuals in question notified about the data collection?

No they were not.

Did the individuals in question consent to the collection and use of their data?

We consider the figures represented in our dataset as camera-facing public (and often historical) figures, with these photos generally taken in public settings per Wikimedia guidelines. However, these individuals did not explicitly consent to the collection of this dataset.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

We will provide a mechanism for identities present in the dataset to request their removal and to update copies of our dataset in distribution accordingly.

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?

No it has not.

Any other comments?

Data Preprocessing

What preprocessing/cleaning was done? (*e.g.*, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instancess, processing of missing values)?

Each image was aligned following the alignment procedure used in the FFHQ dataset, and resized to 256×256 pixels.

Was the "raw" data saved in addition to the preprocessed/cleaned data? (*e.g.*, to support unanticipated future uses)

We saved the url from which the image was collected, and therefore, we could potentially obtain the "raw" images.

Is the preprocessing software available?

The alignment and padding codes are from the FFHQ dataset (code is publicly available).

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

As described in the paper, this dataset has been used to transform faces across time.

Is there a repository that links to any or all papers or systems that use the dataset?

Papers using this dataset will be specified on *Faces Through Time*'s website.

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

This dataset could allow for tasks related to analysis across time as well.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/-cleaned/labeled that might impact future uses?

We made various compositional decisions considering future uses as we discuss in the main paper. For example, we release only one face per identity to prevent utility for facial recognition.

Are there tasks for which the dataset should not be used?

The dataset should not be used for facial recognition applications.

Any other comments?

Data Distribution

Will the dataset be distributed to third parties outside of the entity (*e.g.*, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description

Yes. Researchers at academic institutions (and no others) will be able to request access to the dataset. Requests must specify intended use and discuss ethical considerations of tasks. We place these restrictions to minimize potential for misuse.

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

We will provide researchers whose requests are approved with specific access to download the dataset via email. The dataset will not be publicly available on any website.

When will the dataset be distributed?

Starting October 2022.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

We provide a terms of use agreement with the dataset. The dataset as a whole will be distributed under a non-commercial license and specific images will carry their own licenses (which we also include in the data).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No fees. There will be access restrictions as mentioned above.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Unknown.

Any other comments?

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

The authors of this paper are maintainers of this dataset.

How can the owner/curator/manager of the dataset be contacted (*e.g.*, email address)?

The email addresses of authors are provided in the paper.

Is there an erratum?

At this time, we are not aware of errors in our dataset. However, we will create an erratum as errors are identified.

Will the dataset be updated? If so, how often and by whom? Unknown How will updates be communicated? (*e.g.*, mailing list, GitHub)

The dataset will be updated by the authors on an at-will basis (but no more than once a month) via email to those with access. By terms of use, users will be expected to apply updates before any further use.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (*e.g.*, were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

No such limits are established.

Will older versions of the dataset continue to be supported/hosted/maintained?

N/A

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?

There will not be a mechanism to build on top of the *Faces Through Time* dataset.

Any other comments?