Motivation

Why was the dataset created?

Families In the Wild (FIW) was created to provide images that can be used to study automatic kinship recognition in the unconstrained setting: settings vary across several characteristics (*e.g.*, pose, illumination, resolution, focus,), demographics (*e.g.*, age, gender, race), appearances (*e.g.*, hairstyle, makeup, clothing), and familial meta-data (*i.e.*, families of different sizes that consist of different relationship types). The dynamic nature of the labels allows for data to be parsed for various tasks and applications. Amongst the possible tasks, is the most popular kinship verification task based on face-pair matching: given a pair of facial images, determine whether or not the images are blood relatives.

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

The initial version of FIW was created by Joseph P. Robinson, Ming Shao, and Yun Fu, whom were researchers at the Northeastern University's SMILE Lab at the time of its initial release in 2017.

Who funded the creation of the dataset?

N/A

Composition

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)

The instances of the dataset are pictures of faces, with one face per image.

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

Yes. Explicit relationships are inherent by directory structure. In other words, each folder contains a family. Within each family-folder are subfolders for each member, as is an adjacency matrix describing relationships between members of a given.

```
– F????/
Family archives (1–F, where F is # families)
  – MID?[?]
  Face of that member ID (1–K, where K # members)
  mid.csv:     Member and relationship information
    * matrix represents relationship
    * Name     First name or reference.
    * Gender   [M]ale or [F]emale
```

For example:

MID	1	2	3	Name	Gender
1	0	4	5	name1	female
2	1	0	1	name2	female
3	5	4	0	name3	male

which is made-up of three family members, one per MID across down the rows. Hence, the columns defies the relationship between the respective row. Specifically, the example sets MID1 related to MID2 via 4->1 (Parent->Child). Of course, the opposite holds, *i.e.*, MID2->MID1 is 1->4 (Child-Parent). This example also has a pair of spouses (*i.e.*, related by marriage, and assumed no direct shared ancestors)– MID1 and MID3 are or were married (or significant others), *i.e.*, MID1->MID3 is 5->5. Finally, MID3 is too a parent of MID2, which is not always the case (*i.e.*, relationships are considered independent of one-another). In other words, an MID could have several spouses, and multiple children. However, a child can have at most a single father and mother. Thus, siblings of the same or different pair of parents exist throughout. Nonetheless, all the metadata required to identify the specifics are contained within.

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

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 is a pair of subjects labeled with the name of the person in the image. Some images contain more than one face. The labeled face is the one containing the central pixel of the image—other faces should be ignored as "background".

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)

The dataset comes with specified train/test splits such that there is no family overlap between sets. Therefore, no subject overlap between folds either. The data organized as a table of pairs (datatable.csv or datatable.pkl). Each item is listed with following values across the columns: 'fid', 'fold', 'label', 'p1', 'p2', 'type'. 'fid' indicates the family ID as referenced in FIW, 'fold' indicates the fold to hold the element out for testing (*i.e.*, with remaining folds making up the training set). 'p1' and 'p1' are sample face 1 and 2, respectfully (*i.e.*, facial pair). 'label; is ground-truth (*i.e.*, 0 if unrelated and 1 if related/ true pair). Finally, 'type indicates the relationship in question (*e.g.*, Mother-Son (MS), Father-Daughter (FD), etc.).

Practitioners train algorithms on the training set and evaluate on the test set in a 5-fold fashion. Final performance results are averaged across folds alongside the respective standard deviation. In other words, there are 5 train/test subsets of the dataset- each fold, leave out its items for testing, train on other K-1 (*i.e.*, 5-1=4) folds, then evaluate on test set unseen with respect to the model trained from other folds. This way, all data is used for testing. Furthermore, different training paradigms are allowed while mantaining frozen test sets. As such, we recommend reporting performance on all folds by using leave-one-out cross validation, performing 5 experiments. That is, in each experiment, 4 subsets should be used as a training set and the 5th subset should be used for testing. At a minimum, we recommend reporting the estimated mean accuracy,

$$\hat{\mu} = \frac{\sum_{k=1}^{K} p_k}{K},$$

where K=5 and p_k is percentage of correct classifications for fold k. Along with the standard error of the mean S_E as

$$\mu = \frac{\hat{\sigma}}{\sqrt{K}},$$

where $\hat{\sigma}$ is the estimate of the standard deviation, defined as

$$\mu = \sqrt{\frac{\sum_{k=1}^{K}(p_k - \hat{\mu})^2}{K1}}.$$

Training Paradigms: As the renowned Labeled Faces in the Wild (LFW), FIW supports three training paradigms for the verification task, with the first two the preferred of FIW creators and organizers, however, we decided to include to avoid having reported results that fall outside the two preferred paradigms. Practitioners must clearly state the paradigm followed for all published results.

Step 1. List-Restricted for Training.: This setting does not allow for the family or subject names to be referenced during training or testing. The only labels provided are the ground-truth boolean tags: whether or not a pair of images consist of faces of relatives, and not the identity of the person or any knowledge about family name. Under this paradigm, determining whether multiple pairs of images in the train/test set that belong to the same person(s) and/or family(s) is non-trivial. Such inferences, however, might be made by performing conventional face verification across all pair-wise combinations (*i.e.*, opposed to comparing FID.MID references). Thus, training pairs made-up of matched and mismatched pairs, one can use image equivalence to cluster faces of the same family and even a finer-level of person ID. Pair-list file datatable_restricted.csv is provided with only the labels allowed in this paradigm (*i.e.*, essentially the master-pairs-list file with only columns ['p1', 'p2', 'label']).

Step 2. List-Unrestricted for Training.: This setting allows referencing to the family, subjects, and particular type of relationship bonding the pairs of matched faces, while also providing this same information for the mismatched (*i.e.*, same number of negatives were generated for each type, however, currently types are just being used for analysis of the different types, though certainly possible to leverage knowledge of specific type during training. The file fidlist.csv lists all the family IDs (FIDs), number of members (*i.e.*, member IDs (MIDs)), along with the total number of faces (*i.e.*, face IDs (faceIDs)) between all mumbers of respective family. Matched and mismatched pairs should be used directly as provided in pair-lists. For instance, when processing fold-1, the FIDs of this fold and all pairs they make-up are held out for testing (*i.e.*, frozen as provided in list). For this, positive_pairs.csv is provided with fields for 'p1', 'p2', and 'fold'. For instance, say training pair is F0001.MID1-F0001.MID3 are true kin of type FD, if the former has 5 facial images and the latter has 3 it is then possible to create

$$\binom{n}{2} = \binom{5+3}{2} = 28$$

. Hence, the practitioner can create arbitrary matching/mismatching pairs to best suit their system. Any pairings created and added to list, whether true or false matches, should only be done on training splits, as the unrestricted paradigm allows training list to be created, but not for performance to be reported. The test data of each fold should remain frozen (*i.e.*, unmodified listing of pairs), which then provides platform for fair comparison of performance ratings. We recommend that experimenters first use the List-Restricted paradigm (*i.e.*, use provided lists for both train and test). Then, evolve to the List-Unrestricted paradigm if it is believed that the prospective system would benefit from training with more pairs, more metadata, pairs spanning best representation, etc.. In other words, if it seems benefit could be had by modifying list of training items then move from List-Restricted (above) to List-Unrestricted (this).

Step 3. Unrestricted for Training.: Inherits all characterizations of List-Unrestricted paradigm, but now outside data is allowed to be added to the training. This paradigm is discouraged, as newer, larger versions of FIW will be continually released and, hence, it will be made sure that no subjects added during training are in anyway a part of testing (*i.e.*, not even a subject related to member of test set). Nonetheless, if this paradigm is followed, extensive reporting on the process, sources, and amounts added during training. Regardless the case, **it should be made clear which of these two paradigms are followed.**

What experiments were initially run on this dataset? Have a summary of those results.

Five experiments were initially performed on FIW. Those experiments were kinship verification and family classification; evaluating the proposed semi-supervised clustering method; evaluating CNNs fine-tuned using FIW on KinWild I & II (i.e., transfer-learning), and measuring human performance on kinship verification with a comparison to top-performing algorithms from previous experiments.

Kinship Verification

SphereFace CNN, which was fine-tuned on FIW data, outperformed other benchmarks with an average accuracy of 69.18% (*i.e.*, +1.31% and +2.11% better than ResNet-22+CF and mDML, respectively). The highest verification score tended to be sibling-types followed by parent-child, and, lastly, grandparent-grandchild. Thus, the larger the generational gap, the harder the problem becomes (*i.e.*, the less discriminate facial cues are for determining KIN or NON-KIN).

Family Classification

Family classification also initially saw the top performance as a fine-tuned CNN. Specifically, the ResNet-22 with a Centerface loss obtained an accuracy of 16.18%. The naive baseline approach (one-vs-rest linear Support Vector Machines (SVMs) on top of deep VGG-Face features) achieved an accuracy of just 3.04%. Then, by replacing the softmax layer to target the number of families (*i.e.*, 564), and fine-tuning on FIW, the top-1 accuracy was improved (*i.e.*, +7.38 to 10.42 percent)

Semi-supervised Clustering

Even on the unlabeled data, the proposed method exceeds the K-means baseline. For LCVQE, the pair-wise constraints make the cluster structure unpredictable, vulnerable to deviate from the true one, and, thus, perform worse than the K-means baseline.

Transfer Learning

The fine tuned CNN (*i.e.*, ResNet + Centerface) performed the best, on average, when considering both leader-boards. Furthermore, was its tendancy to yield minimal variation in type-specific scores. On the KinWild I benchmark, we obtained a 4% improvement upon fine-tuning network on FIW data, which is disjoint from that of KinWild (*i.e.*, from 78.4% to 82.4%). Then, on KinWild II, we improved performance by 5.6% (*i.e.*, from 80.0% to 86.6%).

Human Performance on FIW

Human subjects, on average, did not perform so well at recognizing kinship in face photos (*i.e.*, a mean accuracy of 57.5%). This included 150 participants of various backgrounds. In total, there were 200 pairs of faces. Future studies should reduce the sample size (*i.e.*, stick to the closest relationship types), and increase the number of samples (*i.e.*, human subjects).

Close to one year after the initial experiment, we conducted the same experiments with the same volunteers (135 of the original 150 subjects were available and willing to provide another sample). We used the same data pairs, but changed only two aspects: (1) we shuffled the order to avoid patterns from previous evaluation, which could have indirect impact on human decisions; and (2) we did not provide prior knowledge of the relationship type. The purpose here was to see whether or not aging factors impacted the average decision (*i.e.*, being told it is a grandparent-grandchild pair, and with both faces of children, and the one claimed to be grandparent is noticeably aged or monochrome, then we wanted to see if such evidence was impacting decisions). However, and contrary to our original hypothesis, this was not the case, as nearly the same average score results (*i.e.*, 57.8%).

Any other comments?

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 publicly available news sources.

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

No. The dataset only consists of faces and associated names.

Does the dataset relate to people?

Yes. The dataset contains one or more face images of individuals.

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

While sub-population data was not available at the initial release of the dataset, a subsequent paper2 reports the distribution of images by age, race and gender. Table 2 lists these results. The age, perceived gender and race of each individual in the dataset was collected using Amazon Mechanical Turk, with 3 crowd workers labeling each image. After exact age estimation, the ages were binned into groups of 0-10, 21-40, 41-60 and 60+.

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?

Each image is annotated with the name of the person that appears in the image.

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 dataset does not contain confidential information since all information was scraped from news stories.

Any other comments?

Data Collection Process

How was the data associated with each instance acquired?

The names for each person in the dataset were determined by an operator by looking at the caption associated with the person's photograph.

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

The raw images for this dataset were obtained from the Faces in the Wild database collected by Tamara Berg at Berkeley3. The images in this database were gathered from news articles on the web using software to crawl news articles.

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

The original Faces in the Wild dataset is a sample of pictures of people appearing in the news on the web. Labeled Faces in the Wild is thus also a sample of images of people found on the news on line. While the intention of the dataset is to have a wide range of demographic (e.g. age, race, ethnicity) and image (e.g. pose, illumination, lighting) characteristics, there are many groups that have few instances (e.g. only 1.57% of the dataset consists of individuals under 20 years old).

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)?

Students of NEU.

Over what time-frame was the data collected?

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

Unknown

Does the dataset relate to people?

Yes. Each instance is an image of 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 crawled from public web sources.

Were the individuals in question notified about the data collection?

Unknown

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

No. All subjects in the dataset appeared in news sources so the images that we used along with the captions are already public.

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

No. The data was crawled from public web sources, and the individuals appeared in news stories. But there was no explicit informing of these individuals that their images were being assembled into a dataset.

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?

Unknown

Any other comments?

d

How was the data collected? (*e.g.*, hardware apparatus/sensor, manual human curating, software program, software interface/API)

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

Several student volunteers, along with experts overseeing many of the details. No monetary cost was acquired for this effort.

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame of the instances?

FIW was created over a span of months. The time-frame of data (*i.e.*, approximately when the images were captured) does not match that of its creation– imagery was scraped from the web and varies in date, which is metadata we do not have (*i.e.*, no evidence or time stamp indicating its originality).

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 population? What was the sampling strategy (*e.g.*, deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (*e.g.*, geographic coverage)? If not, why not (*e.g.*, to cover a more diverse range of instances)? How does this affect possible uses?

Is there information missing from the dataset and why? (this does not include intentionally dropped instances; it might include, *e.g.*, redacted text, withheld documents) Is this data missing because it was unavailable?

An abundance of faces have been added, but are not released publicly at this moment. The reason is directly related to the lack of public benchmark with lists including the newest face data. Nonetheless, the added data can be provided upon request; also, updated benchmarks will be released as a part of future work.

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)?

The following steps were taken to process the data:

1. *Gathering raw images* First the raw images for this dataset were obtained from the Faces in the Wild dataset consisting of images and associated captions gathered from news articles found on the web. 2. Running the Viola-Jones face detector4 The OpenCV version 1.0.0 release 1 implementation of Viola-Jones face detector was used to detect faces in each of these images, using the function cvHaarDetectObjects, with the provided Haar classifier—cascadehaarcascadefrontalfacedefault.xml. The scale factor was set to 1.2, min neighbors was set to 2, and the flag was set to CV HAAR DO CANNY PRUNING. 3. Manually eliminating false positives: If a face was detected and the specified region was determined not to be a face (by the operator), or the name of the person with the detected face could not be identified (using step 5 below), the face was omitted from the dataset. 4. Eliminating duplicate images: If images were determined to have a common original source photograph, they are defined to be duplicates of each other. An attempt was made to remove all duplicates but a very small number (that were not initially found) might still exist in the dataset. The number of remaining duplicates should be small enough so as not to significantly impact training/testing. The dataset contains distinct images that are not defined to be duplicates but are extremely similar. For example, there are pictures of celebrities that appear to be taken almost at the same time by different photographers from slightly different angles. These images were not removed. 5. Labeling (naming) the detected people: The name associated with each person was extracted from the associated

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

Is the preprocessing software available?

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?

Data Distribution

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

When will the dataset be released/first distributed?

What license (if any) is it distributed under? Are there any copyrights on the data?

Are there any fees or access/export restrictions?

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

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

Is there an erratum?

If the dataset becomes obsolete how will this be communicated?

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

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?

Uses

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

Family classification, large scale search & retrieval, fine-grained categorization, tri-subject verification, hierarchical clustering, to name a few.

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

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

Papers using this dataset and the specified evaluation protocol are listed in http://vis-www.cs.umass.edu/lfw/results.html

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

Papers using this dataset and the specified training/evaluation protocols are listed under “Methods” section of http://vis-www.cs.umass.edu/lfw/results.html

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

The LFW dataset can be used for the face identification problem. Some researchers have developed protocols to use the images in the LFW dataset for face identification.5

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

There is minimal risk for harm: the data was already public.

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

Any other comments?

Benchmarks

CMC

Table 1: Number of pairs used in FIW verification benchmark.

Type	No. Pairs
bb	220,550
ss	220,550
sibs	79,600
fd	99,692
fs	138,177
md	93,347
ms	129,797
gfgd	8,132
gfgs	6,700
gmgd	6,906
gmgs	5,872
Total	100,9323

Legal & Ethical Considerations

If the dataset relates to people (*e.g.*, their attributes) or was generated by people, were they informed about the data collection? (*e.g.*, datasets that collect writing, photos, interactions, transactions, etc.)

If it relates to people, were they told what the dataset would be used for and did they consent? If so, how? Were they provided with any mechanism to revoke their consent in the future or for certain uses?

If it relates to people, could this dataset expose people to harm or legal action? (*e.g.*, financial social or otherwise) What was done to mitigate or reduce the potential for harm?

If it relates to people, does it unfairly advantage or disadvantage a particular social group? In what ways? How was this mitigated?

If it relates to people, were they provided with privacy guarantees? If so, what guarantees and how are these ensured?

Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Does it comply with any other standards, such as the US Equal Employment Opportunity Act?

Does the dataset contain information that might be considered sensitive or confidential? (*e.g.*, personally identifying information)

Does the dataset contain information that might be considered inappropriate or offensive? No. The dataset only consists of faces and associated names.