# **Labelled Faces in the Wild (LFW) Dataset**

Over 13,000 images of faces collected from the web

## **About Dataset**

### **Context**

Labeled Faces in the Wild (LFW) is a database of face photographs designed for studying the problem of unconstrained face recognition. This database was created and maintained by researchers at the University of Massachusetts, Amherst (specific references are in Acknowledgments section). 13,233 images of 5,749 people were detected and centered by the Viola Jones face detector and collected from the web. 1,680 of the people pictured have two or more distinct photos in the dataset. The original database contains four different sets of LFW images and also three different types of "aligned" images. According to the researchers, deep-funneled images produced superior results for most face verification algorithms compared to the other image types. Hence, the dataset uploaded here is the deep-funneled version.

### **Content**

There are 11 files in this dataset. ****lfw-deepfunneled.zip**** is the file containing the images. ****All other 10 files are relevant metadata**** that may help you in forming your training and testing sets for your model. There are two sections below to help you navigate the files better. The first section provides information specifically pertaining to the images. The second section explains the content of each metadata file.

****Image information:****

* ***Image file format***: Each image is available as "lfw/name/name\_xxxx.jpg" where "xxxx" is the image number padded to four characters with leading zeroes. For example, the 10th George\_W\_Bush image can be found as "lfw/George\_W\_Bush/George\_W\_Bush\_0010.jpg"
* ***Image dimensions***: Each image is a 250x250 jpg, detected and centered using the openCV implementation of Viola-Jones face detector. The cropping region returned by the detector was then automatically enlarged by a factor of 2.2 in each dimension to capture more of the head and then scaled to a uniform size.

****Metadata information:****

* ***lfwallnames.csv***: Contains all names of each face in the dataset along with number of images each face has.
* ***lfwreadme.csv***: Comprehensive readme file found on the original database. If there is any information you are missing here or are looking for additional resources you will probably find it in this file. It explains how each .csv file comes into play when forming training and testing models, as well as column metadata information for figuring out what the .csv is talking about. The original website also gives recommendations on training/testing splits and comparison benchmarks.

There are ****two recommended configurations**** for developing training and testing sets (pairs vs people). Depending on which route you choose, you will use the following .csv files:

* ***pairs.csv***: Contains randomly generated splits for 10-fold cross validation specifically for pairs. Use this for the image restricted configuration when forming training sets (refer to readme). There are 10 total sets; 5 sets contain 300 matched pairs, the other 5 sets contain 300 mismatched pairs.
* ***people.csv***: Contains randomly generated splits for 10-fold cross validation specifically for individual faces. Use this for the unrestricted configuration when forming training sets (refer to readme). There are 10 total sets, each with a different amount of people; Set 1: 601. Set 2: 555. Set 3: 552. Set 4: 560. Set 5: 567. Set 6: 527. Set 7: 597. Set 8: 601. Set 9: 580. Set 10: 609.
* ***matchpairsDevTest.csv***: Use this testing set if you decide to go with the pairs configuration. Contains 500 matched pairs of faces for testing set.
* ***matchpairsDevTrain.csv***: Use this training set if you decide to go with the pairs configuration. Contains 1100 matched pairs of faces for training set.
* ***mismatchpairsDevTest.csv***: Use this testing set if you decide to go with the pairs configuration. Contains 500 mismatched pairs of faces for testing set.
* ***mismatchpairsDevTrain.csv***: Use this training set f you decide to go with the pairs configuration. Contains 1100 mismatched pairs of faces for training set.
* ***peopleDevTest.csv***: Use this testing test if you decide to go with the people configuration. Contains 1711 people and 3708 images.
* ***peopleDevTrain.csv***: Use this training set if you decide to go with the people configuration. Contains 4038 people and 9525 images.

### **Acknowledgements**

All data and metadata were originally found on [http://vis-www.cs.umass.edu/lfw/](http://vis-www.cs.umass.edu/lfw/" \t "/home/haris-subrata/Documents\\x/_blank). Please visit the site for other data versions including original, non-aligned data as well as more information on errata and training/testing model resources.

A big thank you and kudos to the creators of this dataset and relevant research:

*Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.*

Specifically for the deep-funneled version of the image data:

*Gary B. Huang, Marwan Mattar, Honglak Lee, and Erik Learned-Miller. Learning to Align from Scratch. Advances in Neural Information Processing Systems (NIPS), 2012.*

Banner photo by John Bakator on Unsplash

### **Inspiration**

* Can you form a model that correctly identifies images that are of the same person?
* What about recognising gender, male or female?
* What pictures are of Al Gore? Are there any faces that look similar to his?