Draft: Evaluation Guidelines for Gender Classification and Age Estimation

Tobias Gehrig, Matthias Steiner, Hazım Kemal Ekenel
{tobias.gehrig, ekenel}@kit.edu

July 1, 2011

1 Introduction

In previous research on gender classification and age estimation did not use a standardised evaluation procedure. This makes comparison the different approaches difficult.

Thus we propose here a benchmarking and evaluation protocol for gender classification as well as age estimation to set a common ground for future research in these two areas.

The evaluations are designed such that there is one scenario under controlled labratory conditions and one under uncontrolled real life conditions.

The datasets were selected with the criteria of being publicly available for research purposes.

File lists for the folds corresponding to the individual benchmarking protocols will be provided over our website at http://face.cs.kit.edu/befit. We will provide two kinds of folds for each of the tasks and conditions: one set of folds using the whole dataset and one set of folds using a reduced dataset, which is approximately balanced in terms of age, gender and ethnicity.

2 Gender Classification

In this task the goal is to determine the gender of the persons depicted in the individual images.

2.1 Data

In previous works one of the most commonly used databases is the Feret database [1, 2]. We decided here not to take this database, because of its low number of images.

2.1.1 Controlled Condition

We propose to use the MORPH-II database [3] for the controlled labratory condition for gender classification.

This database is composed of 55285 color images of 13660 subjects of the age between 16 and 99 years, where 46767 images correspond to male persons and 8518 to female persons. 42671 of these images depict black faces, 10639 white, 1753 hispanic, 160 asian, 57 indian and 5 faces are of other ethnicities. The images have varying resolutions of either 200×240 or 400×480 pixels.

While this dataset is highly imbalanced towards black male persons, this adds an additional challenge, which could also point out the generalizability of machine learning approaches.

The dataset can be requested from:

http://www.faceaginggroup.com/projects-morph.html.

2.1.2 Uncontrolled Condition

For the uncontrolled condition we decided to use the Labeled Faces in the Wild (LFW) [4] dataset, which contains faces of 5749 individuals (4263 male, 1486 female) collected from the web using a Viola-Jones face detector. Of these there are 1680 people for which more than one image is available. This results in 10256 male images and 2977 female images. These color images have an resolution of 250×250.

The dataset can be requested from:

http://vis-www.cs.umass.edu/lfw/.

2.1.3 Data Format

File lists for the folds corresponding to the individual benchmarking protocols will be provided over our website at http://face.cs.kit.edu/befit. These files are structured such that there is one line per image. Each line starts with the filename followed by the fold ID and the gender. Each of the values is separated by a tab. The gender can be either `M` for male or `F` for female.

An example line looks like the following:

```
224896_00M25.JPG        0        M
```

where `224896_00M25.JPG` is the filename of the image, 0 is the fold ID, and `M` shows that the person depicted on the image is of male gender.

2.2 Evaluation Metrics

A commonly used metric is the *accuracy*:

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

where TP is the number of correctly classified positive samples, FP the number of samples that have been classified incorrectly as positives, TN the number of correctly classified negative samples, and FN the number of samples that have been classified incorrectly as negatives. This metric specifies the overall accuracy of the classification, but has several disadvantages. One is that a classifier that has unbalanced accuracies between male and female images can lead to a higher value of accuracy than a classifier that has more balanced accuracies between the genders, which is what we are more interested in. It also is highly influenced by how balanced the dataset is. Thus we propose to use it only for historic reasons. Additionaly, we suggest to use the *true positive rate*

$$TPR = \frac{TP}{P}, \tag{2}$$

where P is the total number of positive samples, and the *true negative rate*

$$TNR = \frac{TN}{N}, \tag{3}$$

where N is the total number of negative samples, to determine seperately the classification accuracy for male and female respectively. Based on this we get also a slightly more robust variant of the accuracy in terms of dataset balance, the *average correct rate* (ACR):

$$ACR = \frac{TPR + TNR}{2} \tag{4}$$

Finally, as a single valued metric, that is more robust to imbalanced datasets or imbalanced classification, we propose to use the *area under the receiver operator characteristic curve* (AUC) as was also suggested by Martín-Félez et al. [5] for gait-based gender classification. This metric returns its maximum value (1), when we have an optimal system, which has a TPR of 1 and a *false positive rate* (FPR) of 0. Since in this case the *receiver operator characteristic* (ROC) curve directly jumps from the lower left ($TPR = 0$, $FPR = 0$) to the upper left ($TPR = 1$, $FPR = 0$) and then stays at $TPR = 1$. In case of a non-optimal system we prefer a system that has a balanced accuracy. An algorithm to calculate the AUC was given by Fawcett in [6]. An upper-bound on its uncertainty is given by [7]:

$$s = \sqrt{\frac{AUC(1 - AUC)}{\min\{P, N\}}} \tag{5}$$

2.3 Benchmarking Protocol

For evaluating gender classification approaches 5-fold cross-validation shall be used. To prevent algorithms from learning the identity of the persons in the training set rather than the gender it has to be made sure that all images of individual subjects are only in one fold at a time. Additionally, the folds are selected in such a way that the distribution of age, gender and ethnicity in the

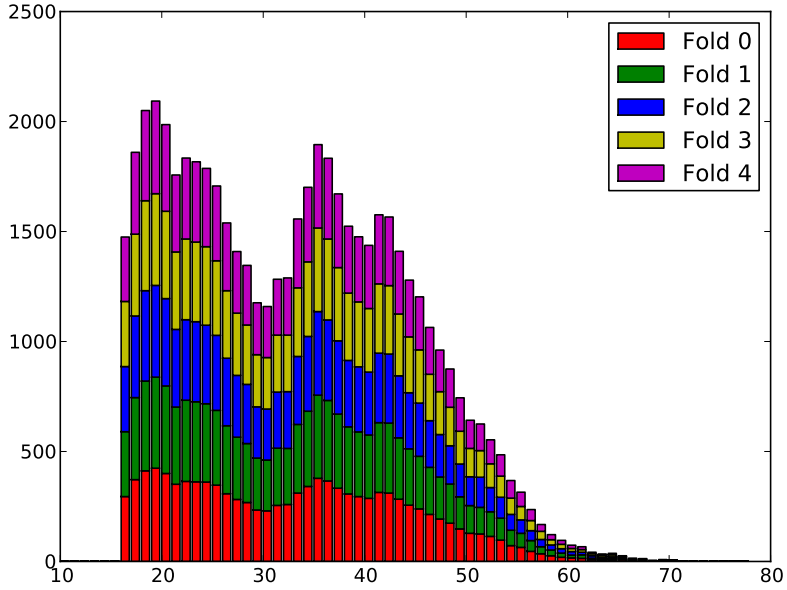

Figure 1: Distribution of the MORPH-II database images over age in the individual folds.

folds is similar to the distribution in the whole database. In Fig. 1, Fig. 2 and Fig. 3 the distribution of the images over age, gender and ethnicity respectively in the individual folds is depicted for the MORPH-II database. In Fig. 4 and Fig. 5 the distribution of the images over gender and ethnicity respectively in the individual folds is depicted for the LFW database.

3 Age Estimation

In this task the goal is to estimate the age of the persons depicted in the individual images.

3.1 Data

A challenging problem in choosing a dataset for age estimation is the need for a large age range which should ideally be equally distributed over all ages.

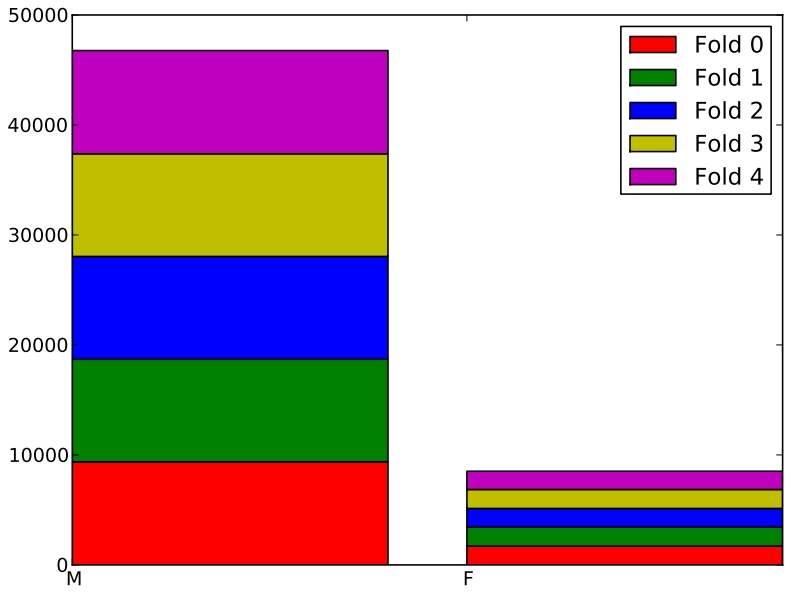

Figure 2: Distribution of the MORPH-II database images over genders in the individual folds.

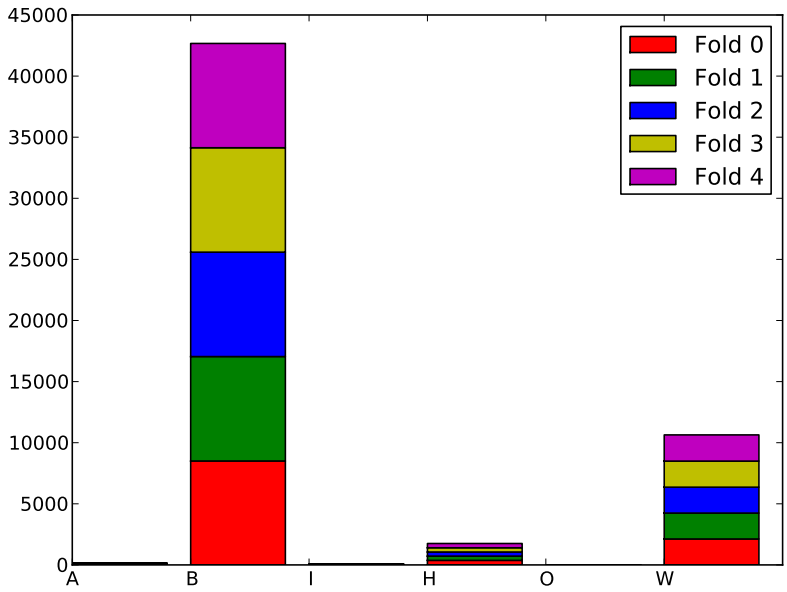

Figure 3: Distribution of the MORPH-II database images over the ethnicity in the individual folds.

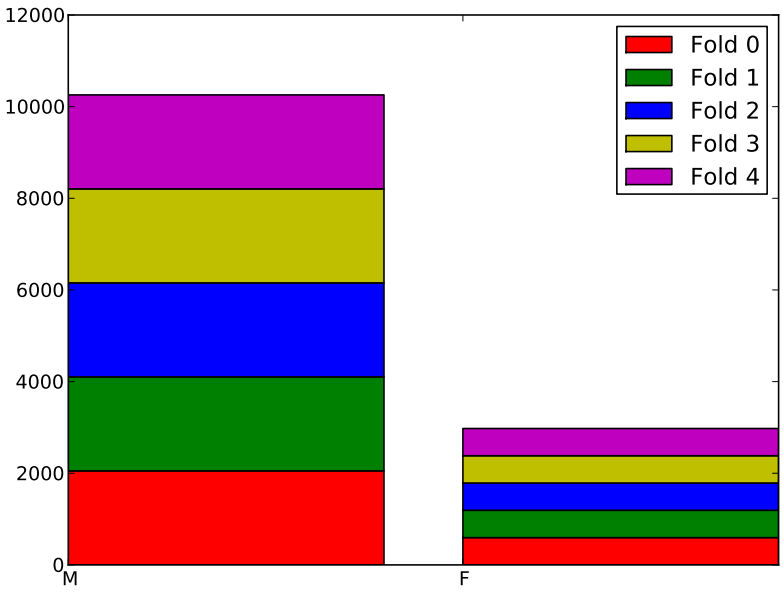

Figure 4: Distribution of the LFW database images over genders in the individual folds.

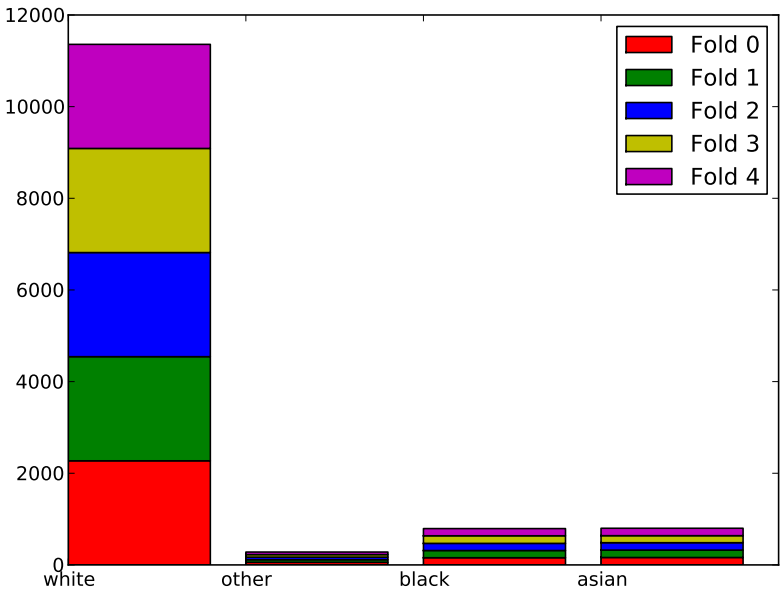

Figure 5: Distribution of the LFW database images over the ethnicity in the individual folds.

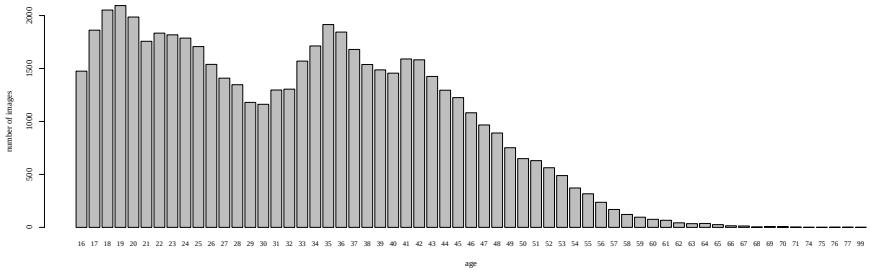

Figure 6: Distribution of the MORPH-II database images over ages.

3.1.1 Controlled Condition

We propose to use the MORPH-II database [3] for the controlled labratory condition also for age estimation.

This database is composed of 55608 color images of 13673 subjects of the age between 16 and 99 years, where 47057 images correspond to male persons and 8551 to female persons. 42897 of these images depict black faces, 10736 white, 1753 hispanic, 160 asian, 57 indian and 5 faces are of other ethnicities. The images have varying resolutions of either 200×240 or 400×480 pixels. The distribution of the images over ages is depicted in Fig. 6.

While this dataset is highly imbalanced towards black male persons and missing images of persons below the age of 16, this adds an additional challenge, which could also point out the generalizability of machine learning approaches.

The dataset can be requested from:

http://www.faceaginggroup.com/projects-morph.html.

3.1.2 Uncontrolled Condition

The Face and Gesture Recognition Research Network (FG-NET) aging database [8] is proposed to be used for the uncontrolled real-life condition. The database contains on average 12 pictures of varying ages between 0 and 69, for each of its 82 subjects. Altogether there are a mixture of 1002 color and greyscale images, which were taken in totally uncontrolled environments. Each was manually annotated with 68 landmark points. In addition there is a data file for every image, containing type, quality, size of the image and information about the subject such as age, gender, spectacles, hat, mustache, beard and pose.

One particular problem with this dataset is the fact, that images are not equally distributed over age and thus only few images of persons older than 40 are available.

The dataset can be requested from:

http://www.fgnet.rsunit.com/.

3.1.3 Data Format

File lists for the folds corresponding to the individual benchmarking protocols will be provided over our website at `http://face.cs.kit.edu/befit`. These files are structured such that there is one line per image. Each line starts with the filename followed by the fold ID and the age in years. Each of the values is separated by a tab.

An example line looks like the following:

```
224896_00M25.JPG        0        25
```

where `224896_00M25.JPG` is the filename of the image, 0 is the fold ID, and 25 shows that the age of the person depicted on the image is 25 years.

3.2 Evaluation Metrics

A commonly used metric for age estimation is the *mean absolute error* (MAE) defined by:

$$MAE = \frac{\sum_{i=1}^{N_t} |\hat{a}_i - a_i|}{N_t}, \tag{6}$$

where $\hat{a}_i$ is the estimated age and a_i the ground truth for the i-th of N_t test samples [9].

Another measure that is commonly calculated to evaluate the performance is the *cumulative score* (CS):

$$CS(\theta) = \frac{N_{e \leq \theta}}{N_t} \times 100\%, \tag{7}$$

where $N_{e \leq \theta}$ is the number of estimates that have an absolute error below or equal to θ [9].

Since the MAE doesn't give any hint on the performance over the different ages recent publications also specified the *MAEs per decade* (MAE/D), where the MAE is calculated for each decade seperately [9].

Additionally, to see the confusion of the age estimation over these decades, we suggest to also provide a confusion matrix over the decades.

Finally, since we still would like to have single valued metric and MAE is sensitive to imbalanced datasets and imbalanced performance, we propose in addition to use the *average over MAEs per year* (AMAE/y):

$$AMAEy = \frac{\sum_{i=1}^{N} MAE_{a_i}}{N}, \tag{8}$$

where a_i is the i-th age class, N the number of age classes and MAE_{a_i} is the MAE for age a_i.

3.3 Benchmarking Protocol

Due to the size of the MORPH-II database the 5-fold cross-validation evaluation scheme shall be used for the evaluation of the approaches for the controlled condition. To prevent algorithms from learning the identity of the persons in the training set rather than the age it has to be made sure that all images of individual subjects are only in one fold at a time. Additionally, the folds are selected in such a way that the distribution of age, gender and ethnicity in the folds is similar to the distribution in the whole database. In Fig. 1, Fig. 2 and Fig. 3 the distribution of the images over age, gender and ethinicity respectively in the individual folds is depicted.

For the evaluation of the approaches for the uncontrolled condition the *leave-one-person-out* (LOPO) evaluation scheme shall be used, since this is a common evaluation scheme for this database, due to the small size of the FG-NET Aging database. For LOPO, all samples of a single person are used for testing and the remaining samples for training. This is done for all subjects so that each person is once used for testing, resulting in 82 folds. This evaluation scheme makes sure that images of a person are not in the testing and training set at the same time, so that the classifier cannot learn some "intra personal" relations.

References

[1] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, 1998.

[2] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[3] K. Ricanek Jr. and T. Tesafaye, "MORPH: A Longitudinal Image Database of Normal Adult Age-Progression," in *IEEE 7th International Conference on Automatic Face and Gesture Recognition (FGR'06)*, Southampton, UK, Apr. 2006, pp. 341–345.

[4] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," University of Massachusetts, Amherst, Tech. Rep., Oct. 2007.

[5] R. Martín-Félez, R. A. Mollineda, and J. S. Sánchez, "A Gender Recognition Experiment on the CASIA Gait Database Dealing with its Imbalanced Nature," in *International Conference on Computer Vision Theory and Applications (VISAPP 2010)*, 2010, pp. 439–444.

[6] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[7] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, "Toward practical smile detection." *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 11, pp. 2106–11, Nov. 2009.

[8] "FG-NET Aging Database." [Online]. Available: http://www.fgnet.rsunit.com/

[9] K. Luu, T. Bui, C. Suen, and K. Ricanek, "Spectral Regression based age determination," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2010)*, no. 1993. IEEE, 2010, pp. 103–107.