Gender Recognition Using Cognitive Modeling

Jens Fagertun, Tobias Andersen, and Rasmus Reinhold Paulsen

Technical University of Denmark
Department of Informatics and Mathematical Modelling
Lyngby, Denmark

Abstract. In this work, we use cognitive modeling to estimate the *"gender strength"* of frontal faces, a continuous class variable, superseding the traditional binary class labeling. To incorporate this continuous variable we suggest a novel linear gender classification algorithm, the Gender Strength Regression. In addition, we use the gender strength to construct a smaller but refined training set, by identifying and removing ill-defined training examples. We use this refined training set to improve the performance of known classification algorithms. Also the human performance of known data sets is reported, and surprisingly it seems to be quite a hard task for humans. Finally our results are reproduced on a data set of above 40,000 public Danish LinkedIN profile pictures.

Keywords: Gender recognition, Linear Discriminant Analysis, Support Vector Machines, Cognitive Modeling, Linear Regression.

1 Introduction

Determining gender is a natural, subconscious task performed everyday in human interaction. By continuously training of this task on a daily basis, the human mind is shaped to become a gender recognition expert. Despite the strong belief that this is the case, there are very few scientific studies to support this. The studies that make a comparison between human vs. machine gender recognition performance are nearly two decades old, such as [1]. However, it is probably safe to assume that we humans do perform well as gender recognition experts, even with limited scientific documentation of the subject. This gives rise to some interesting questions: "Can we understand how humans perform the task of gender recognition?" and "Can we use this information to enhance machine learning models?".

It should be noted, that when confined only to non-intrusive data (such as visual or audio data) the two populations (male and female) cannot be perfectly separated. There are border-line cases where determining gender is also extremely hard for humans; cases where sports athletes need to undergo a DNA test to exactly determine the gender, come to mind.

The problem of automatic gender classification has been studied for years. Many algorithms have been proposed that employ different methods, including Linear Discriminant Analysis (LDA) [2], Haar-like wavelets [3], Locally Binary

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part II, LNCS 7584, pp. 300–308, 2012.

Patterns (LBP) [3] and Support Vector Machines (SVM) [4]. All these methods work on a binary understanding of class association, which does not necessarily enable the method to exclude ill-defined samples that could lead to a performance decrease. This work tries to address this shortcoming by analyzing how humans perform gender classification and incorporating this information into both a novel and well-known pattern recognition algorithms.

Obviously, humans perform gender recognition based on a variety of data, such as face, hair style, clothes, height, voice and movement patterns. In this article we look exclusively at faces. To address the previously posed questions, this work will look into the human performance. We estimate a continuous *"gender strength variable"* and use it to construct a novel linear gender classification algorithm, the Gender Strength Regression (GSR). In addition, we look into using this gender strength variable to construct a smaller but refined training set, by identifying and removing ill-defined training examples. We then assess if this refined training set improves performance of known classification algorithms.

The structure of the paper is as follows. Section 2 introduces the data sets used in this work. Section 3 describes the cognitive test that estimates the gender strength variable. Section 4 presents the Gender Strength Regression method. Section 5 presents the experiments and the obtained results. Finally, Section 6 presents a discussion and conclusion on this work.

2 Data Sets

In this work we use four data sets in order to ensure that our results are not biased towards a single protocol of acquiring facial images. The data sets are subsets of publicly available data, filtered so that they only contain frontal images with no extreme lighting or extreme facial expressions. The images have been cropped and rotated to only show the central face, excluding outer cheeks, lower chin and hair. Every image has been converted to gray scale and standardized by a histogram equalization. Furthermore, the data sets include the same number of males and females. This should motivate the observer to employ a response criterion that is not biased towards one of the two response categories. The data sets used are described in Table 1 with random samples shown in Figure 1.

3 Cognitive Test

The object of the cognitive test is to estimate the gender strength of each sample. The gender strength variable is defined as a variable in the interval [0 1] where close to zero signifies belonging poorly to the sample's group, whereas one signifies belonging strongly to the sample's group. To encode the gender in the variable, females will be denoted in the negative interval [-1 0] and males in the positive interval [0 1].

In order to estimate the gender strength variable, two cognitive user tests were devised and conducted on four and six test participants, respectively. The gender strength variable is estimated as a latent variable encoded by the time it takes participants to perform the gender classification task.

Table 1. Data sets used in this work

Data set	Total number of images used
AR [6]	298
XM2VTS [7]	552
FERET [8]	836
Public Danish LinkedIN profile pictures	200

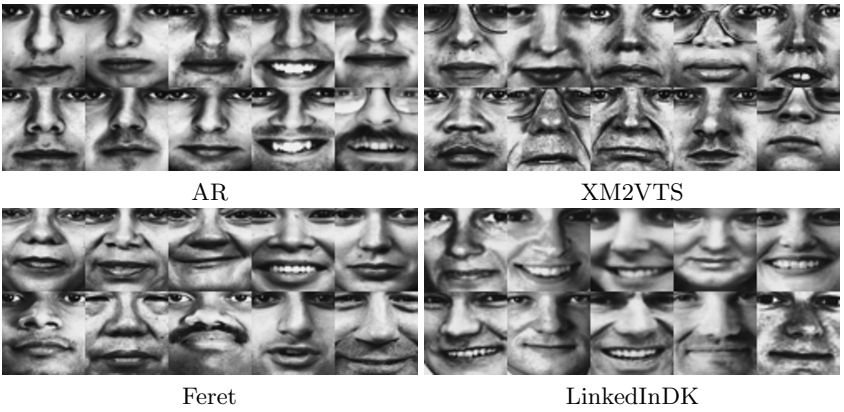

Fig. 1. Random sample images from the four data sets, with females and males in the first and second row, respectively

The cognitive test is conducted with the psychtoolbox for matlab [9], to ensure precise exposure/answer times. In the first test (time limited) every facial image is shown for precisely 200 ms, followed by a scrambled image of the original image (also shown for 200 ms) to disrupt the after image, which is sustained neural activity in the visual system following a visual stimulus. In the second test (not time limited) every facial image is shown until the user makes the classification. A graphical illustration of the tests is shown in Figure 2.

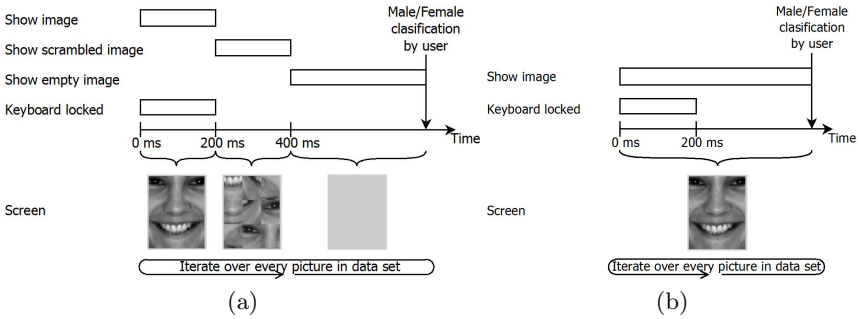

Fig. 2. Timeline for the cognitive tests, (a) time limit and (b) no time limit. There is a 550 ms delay from a user classification to the next picture is shown.

The gender strength variable is estimated by standardizing the test persons answer times, so that all answer times from one test person (per data set) has mean zero and unit variance. Then, a pooling of the standardized answer times from all test persons is performed. Finally, the pooled answer times below the mean are encoded to one, the outliers (above three standard deviations) are encoded to zero and the remaining are linearly encoded between the values zero and one.

4 Gender Strength Regression

In classification problems, training data is usually only labeled with which class it belongs to. As a result, classifiers, such as LDA and SVM, have only a binary understanding of class membership, which they use in the objective function to construct the classification method. LDA seeks to minimize the within class variance while maximizing the between class variance, whereas SVM tries to maximize the margin (distance) between classes. Due to this binary understanding of class, these methods work on the complete data set. The performance of methods such as LDA and SVM depends solely on the given training data, where more training data does not necessarily yield a better performance.

In this paper we suggest enhancing the performance of classifiers such as LDA and SVM by removing ill-defined training samples. In this work, ill-defined samples are defined as cases where more than 50% of the test participants in the the human cognitive tests fail to perform a correct classification.

Furthermore, we suggest a new classifier that seeks to estimate the gender strength variable by regression. A linear least squares regression has the objective function

$$\min_{\beta} \|\beta \mathbf{X} - \mathbf{y}\|^2, \tag{1}$$

with the closed form solution

$$\beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}, \tag{2}$$

where $\mathbf{X}$ is a matrix with samples and variables in the rows and columns, respectively. $\mathbf{y}$ is a vector with the gender strength scores, also called the response variable. β is the coefficient for the regression line. In an optimal system ((1) is small) the decision threshold would be zero in accordance with how we encoded the response variable in the regression, see Section 3. However, this will not always be the case. A more accurate decision threshold can be calculated by

$$\frac{\sum_{i=1}^{k} \beta \mathbf{x}_i}{2k} + \frac{\sum_{j=1}^{l} \beta \mathbf{x}_j}{2l}, \tag{3}$$

where i and j belong to samples of class male and female, respectively. k and l are the sizes of the two classes. This assumes that the two classes have the same distributions.

5 Experiments

5.1 Cognitive Test

In the cognitive test, the False Classification Rate (FCR) is reported in Table 2. It can be seen that in the first test where participants only saw the central face for a fraction of a second the performance was not bad, in the sense that it is clearly better than guessing (test participants reported that they felt like they were guessing most of the time). When the time constraint was removed, the performance increased. Also it seems that the individuals in the AR and Feret data sets had a bias toward masculinity, resulting in more females that were wrongly classified as a male than compared to the two other data sets.

Table 2. False classification rates for the cognitive test with standard deviations

Data set		Test with time constraint	Test without time constraint	Combined
AR	Male*	0.068 ± 0.047	0.006 ± 0.008	0.031 ± 0.042
	Female**	0.107 ± 0.062	0.164 ± 0.070	0.141 ± 0.070
	Total	**0.174 ± 0.096**	**0.169 ± 0.066**	**0.171 ± 0.074**
XM2VTS	Male*	0.084 ± 0.054	0.035 ± 0.015	0.055 ± 0.042
	Female**	0.097 ± 0.045	0.053 ± 0.030	0.071 ± 0.041
	Total	**0.181 ± 0.050**	**0.088 ± 0.025**	**0.126 ± 0.059**
Feret	Male*	0.048 ± 0.052	0.027 ± 0.009	0.035 ± 0.033
	Female**	0.163 ± 0.131	0.079 ± 0.039	0.112 ± 0.092
	Total	**0.211 ± 0.121**	**0.106 ± 0.043**	**0.148 ± 0.094**
LinkedInDK	Male*	0.059 ± 0.038	0.038 ± 0.020	0.046 ± 0.029
	Female**	0.104 ± 0.048	0.028 ± 0.019	0.058 ± 0.050
	Total	**0.163 ± 0.053**	**0.065 ± 0.028**	**0.104 ± 0.063**

* Male wrongly classified as a female.
** Female wrongly classified as a male.

5.2 Projection Lines

To obtain an understanding of how GSR classifies gender, all samples in the data sets are projected onto the respective GSR line. Only the time limited test is used as we want to see the first features we as humans use to perform gender classification. The estimated Gaussian distribution of these projections is shown in Figure 3, together with synthesized faces corresponding to points on the regression line.

5.3 Recognition Results

Cross-Validation Evaluation Gender classification was performed for all data sets using GSR and compared to LDA and SVM[3] with a linear kernel in a leave-one-out cross-validation scheme (Train on all data, except one sample. Then

[3] Matlab implementation of LDA and SVM is used.

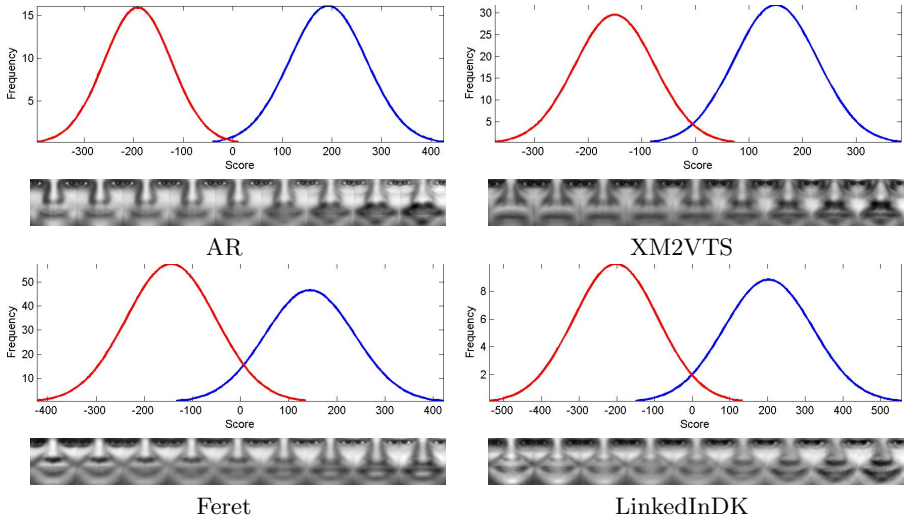

Fig. 3. The estimated gaussian distribution of the data sets projected to the GSR projection line, with synthesized faces corresponding to points on the regression line

test on this "unseen" sample and iterate this process over all samples in the given data set). Three experiments were conducted where the gender strength variable and the ill-defined samples were estimated by (case 1) only the time-limited test, (case 2) only the non-time-limited test and (case 3) by both the time-limited and non-time-limited test combined. The results are presented in Table 3. As the dimension of the variables is much higher than the samples for all data sets, a dimensionality reduction by Principal Component Analysis (PCA) [5] was preformed prior to building the models. The PCA was set to retain 95% of the data set variance.

Generalization Evaluation. To see how the algorithms perform on images not included in four data sets, an experiment was conducted, where all four data sets were collapsed into one training set of 1886 images and these were tested on a data set consisting of 40,692 images from public Danish LinkedIN profiles (LinkedInDKFULL). In this experiment we used a radial basis function kernel for SVM. Finding the optimal parameters for a non-linear SVM is still an open problem, therefore we employ a gridsearch strategy to determine the C and γ, in the grid $[10^{-2}10^{5}]$ and $[10^{-5}10^{1}]$, respectively. It was found that $C = 10^4$ and $\gamma = 10$ was the optimal setup. The results are presented in Table 4, where it can be seen that removing ill-defined training samples had a huge impact.

6 Discussion and Conclusion

In this work we have presented the gender strength variable, a new way of determining class membership that supersedes the conventional binary classes

Table 3. FCR obtained by a leave-one-out cross-validation scheme for case 1 to 3, including (full training set) or excluding (refined training set) ill-defined training examples. Bold signifies lowest FCR for a data set in the corresponding test and green signifies lowest FCR for a data set between tests in (full training set) and (refined training set).

	Full training set				Refined training set			
	Data set	**LDA**	**SVM**	**GSR**	**Data set**	**LDA**	**SVM**	**GSR**
Case 1	AR	0.0604	0.0604	**0.0570**	AR	**0.0537**	**0.0537**	0.0671
	XM2VTS	0.1304	0.1304	**0.1250**	XM2VTS	0.1286	0.1286	**0.1250**
	Feret	**0.1364**	**0.1364**	0.1495	Feret	**0.1280**	0.1292	0.1388
	LinkedInDK	**0.1950**	0.2000	0.2000	LinkedInDK	**0.1800**	**0.1800**	0.2000
Case 2	**Data set**	**LDA**	**SVM**	**GSR**	**Data set**	**LDA**	**SVM**	**GSR**
	AR	**0.0604**	**0.0604**	0.0638	AR	**0.0604**	**0.0604**	0.0839
	XM2VTS	0.1304	0.1304	**0.1232**	XM2VTS	**0.1196**	**0.1196**	0.1304
	Feret	0.1364	0.1364	**0.1304**	Feret	**0.1244**	0.1256	0.1292
	LinkedInDK	**0.1950**	0.2000	0.2050	LinkedInDK	**0.1900**	**0.1900**	0.1950
Case 3	**Data set**	**LDA**	**SVM**	**GSR**	**Data set**	**LDA**	**SVM**	**GSR**
	AR	0.0604	0.0604	**0.0570**	AR	**0.0570**	**0.0570**	0.0638
	XM2VTS	0.1304	0.1304	**0.1268**	XM2VTS	**0.1159**	**0.1159**	0.1232
	Feret	**0.1364**	**0.1364**	0.1376	Feret	0.1232	**0.1184**	0.1352
	LinkedInDK	**0.1950**	0.2000	0.2100	LinkedInDK	**0.1700**	**0.1700**	0.2100

Table 4. FCR obtained by training on all four data sets and testing on 40,692 public Danish LinkedIN profile pictures, including (a) or excluding (b) ill-defined training examples. Bold signifies lowest FCR for a data set in the corresponding test and green signifies lowest FCR for a data set between tests in (a) and (b).

Data set	**LDA**	**SVM**	**GSR**
LinkedInDKFULL	0.3762	**0.3227**	0.3521

(a)

Data set	**LDA**	**SVM**	**GSR**
LinkedInDKFULL	0.2520	**0.2172**	0.2373

(b)

normally used in gender recognition. We have assessed the benefit of the gender strength variable by identifying and removing ill-defined training samples to improve existing methods such as LDA and SVM. Employing the refined training set on LDA and SVM (both in a linear and non-linear version) gave a significant improvement. We could see a performance increase when pooling answers from both cognitive tests (with and without time constraints) probably due to the fact that we obtained a more robust estimation of ill-defined training samples. This indicates that not only the volume of the training data is important but also the quality of the training data. However, how to optimally determine ill-defined training samples is still an open question.

Interestingly we also observed an improvement with SVM. As SVM manly focuses on samples close to the decision boundary one could expect that removing ill-defined samples would lead to a decrease in performance. However, here we should remember that the two populations (male and female) cannot be perfectly separated, and ill-defined samples my not lie adjacent to the decision boundary, which results in the observed performance increase.

A new method was also devised - the Gender Strength Regression. Results indicate that without removing ill-defined training samples GSR performs similarly to or better than the LDA and SVM (in a linear version). However, when using the refined training set, LDA and SVM outperforms GSR. Also the complexity of the data seems to be best modeled by non-linear methods, however these methods need more tuning of parameters.

Human performance was reported, to the authors' knowledge for the first time, on the four data sets. When comparing the performance of human vs. automatic machine learning algorithms, it can be seen that the machine learning algorithms outperform humans on the AR data set. Machine learning algorithms perform only 1-3% worse than humans on the Feret and XM2VTS data sets. But when going from the relatively constrained data sets (similar frontal poses) to the more unconstrained data set of Danish LinkedIN profile pictures, humans outperform the machine learning algorithms. It should be kept in mind that all tests were performed on the central face in gray scale images.

Finally, the results obtained for the four data sets were validated by reproducing the trend that a performance boost was achieved when removing ill-defined training samples, by collapsing the four training sets and testing them on a data set of 40,692 public Danish LinkedIN profile pictures.

It should be noted that the gender strength variable seems to be well estimated by a linear encoding of the gender classification response times, however it is not known if this is the optimal encoding.

References

1. Bruce, V., Burton, A., Hanna, E., Healey, P., Mason, O., Coombes, A., Fright, R., Linney, A.: Sex discrimination: how do we tell the difference between male and female faces? Perception 22(2), 131–152 (1993)
2. Bekios-Calfa, J., Buenaposada, J.M., Baumela, L.: Revisiting Linear Discriminant Techniques in Gender Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(4), 858–864 (2011)
3. Mäkinen, E., Raisamo, R.: Evaluation of Gender Classiffcation Methods with Automatically Detected and Aligned Faces. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(3), 541–547 (2008)
4. Moghaddam, B., Yang, M.-H.: Learning Gender with Support Faces. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(5), 707–711 (2002)
5. Kirby, M., Sirovich, L.: Application of the Karhunen-Loeve procedure for the characterization of human faces. IEEE Transactions on Pattern Analysis and Machine Intelligence 12, 103–108 (1990)
6. Martinez, A.M., Benavente, R.: The AR Face Database. Technical Report, Computer Vision Center Purdue University (1998)

7. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: XM2VTSDB: The Extended M2VTS Database. In: Second International Conference on Audio and Videobased Biometric Person Authentication (1999)
8. Phillips, H., Moon, P., Rizvi, S.: The FERET evaluation methodology for face recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(10) (2000)
9. Brainard, D.H.: The Psychophysics Toolbox. Spatial Vision 10(4), 433–436 (1997)