A COMPARISON OF TECHNIQUES FOR ROBUST GENDER RECOGNITION

R.N. Rojas-Bello, L.F. Lago-Fernández, G. Martínez-Muñoz, M.A. Sánchez-Montañés

Departmento de Ingeniería Informática, Universidad Autónoma de Madrid, Spain

ABSTRACT

Automatic gender classification of face images is an area of growing interest with multiple applications. Appropriate classifiers should be robust against variations such as illumination, scale and orientation that occur in real world applications. This can be achieved by normalizing the images in order to reduce those variations (alignment, re-scaling, histogram-equalization, etc.), or by extracting features from the original images which are invariant respect to those variations. In this work we perform a robust comparison of eight different classifiers across 100 random partitions of a set of frontal face images. Four of them are state-of-the-art methods in automatic gender classification that use image normalization (SVMs, Neural Networks, ADABOOST and PCA+LDA). The other four strategies use invariant features extracted by SIFT (BOW, Evidence Random Trees, NBNN and Voted Nearest-Neighbor). The best strategies are SVM using normalized images and NBNN, the latter having the advantage that no strong image pre-processing is needed.

Index Terms— Automatic gender classification, image analysis, machine learning

1. INTRODUCTION

Automatic gender classification of face images is an area of growing interest with multiple applications such as demographic data collection or facial expression recognition (for a recent review see [1]). In most practical applications, the properties of the images that will be used in the exploitation phase (illumination, pose, scale, orientation, occlusions, etc.) may differ much from those used for training the classification system. Therefore, an appropriate gender classifier should be robust against these variations.

There are two strategies to achieve this robustness: (i) to implement a pre-processing stage that removes as much as possible these variations in the input images; and (ii) to construct classifiers that use invariant features with respect to the mentioned variations. The first strategy is the standard one in gender classification [1]. Common approaches include image normalization, image alignment to reduce scale/orientation variability, and histogram equalization to remove illumination variability. The second approach has been widely explored for image categorization [2], but there are not much published results in the context of gender recognition [3].

In this work we perform a comparison between classification methods using both strategies. Firstly we consider the best state-of-the-art (pixel intensity based) classifiers according to [4] (SVMs, Neural Networks), as well as AdaBoost and PCA+LDA. All these methods use 24×24 pixel aligned, re-scaled and histogram-equalized images of the faces as input. Secondly we explore classifiers based on SIFT invariant features [5]. Due to the invariant properties of the SIFT keypoints, no normalization is needed in this case. We consider Bag Of Word (BOW) based classifiers [2], Evidence Random Trees (ERT) [6] and two different strategies based on nearest neighbors, including Naïve Bayes Nearest Neighbor (NBNN) [7].

The performance of the different classification methods is evaluated using 100 random training/test partitions of a subset of the public Feret database [8] containing only frontal images. This allows us to quantify both the average error and the robustness (given by the standard deviation of the error) of each method, avoiding the possibility that a specific training/test partition favours one of the techniques. Our comparison shows that the SVM trained with normalized images has the highest performance. The best SIFT-based method is the NBNN, with almost the same error rate. Nemenyi's test does not find any significant statistical differences between the two methods, but the NBNN has the additional advantage of not requiring strong image pre-processing.

2. METHODS

In this section we describe the database and the classification techniques we employ.

2.1. Database

We use images from the original Gray Feret database, which is now a subset of the Colour Feret database [8]. We used the same subset as in [4], which consists of 411 frontal images (with only one image per subject), 212 of which belong to class male and 199 to class female. In order to facilitate comparisons, the authors provided a link with the gender, face coordinates and eye positions of all the images, and divided

This work has been supported by CDTI (project INTEGRA) and DGUI-CAM/UAM (project CCG10-UAM/TIC-5864)

them into a training set with 304 images and a test set with 107 images. Here we use the same labels, face and eye coordinates. However, instead of using a single training/test partition, we generated 100 random partitions each consisting of 304 training images and 107 test images. This allows to quantify not only the average error obtained by each classification technique, but also its robustness (given by the standard deviation of the error).

2.2. Classification techniques

We compare eight classification methods that follow two main approaches. Firstly we consider classifiers which are trained using normalized images as input. Secondly we consider classifiers based on SIFT keypoints extracted from the images.

2.2.1. Classification methods using normalized images

A standard approach to gender classification is to train a classifier, such as a neural network or a support vector machine (SVM) [9], using as input the pixel intensity values of the face images. This kind of approach has been shown to achieve high performance rates [4, 1], but a strong preprocessing stage (normalization, alignment, scaling and/or histogram equalization) is needed. Based on the state of the art, we normalize and align the images using the eye positions, and rescale them to a resolution of 24×24 pixels. We use the same face and eye coordinates reported by [4]. Then histogram equalization is applied, and the resulting vectors of 576 pixel intensities are used to feed the classifiers. The following classifiers are analyzed:

Neural networks. We use a network architecture with one hidden layer (two neurons with tansig activation function), and a linear output. The input values are scaled to the range [-0.5, 0.5]. The networks are trained with the standard backpropagation algorithm. We use 25% of the training images as a validation set to avoid overfitting. For each of the 100 random partitions, we trained 20 networks using different, randomly chosen, validation sets, and selected the network with the lowest validation error.

Support Vector Machines. We use the SVM implementation provided by the LIBSVM library [10], taking as input the histogram equalized pixel intensity values scaled to the range [-1, 1]. We used a RBF kernel, and a 10-fold cross validation was performed in order to adjust the complexity parameter and the kernel width.

ADABOOST. We used the Gentle Adaboost implementation provided by the GML Adaboost Matlab Toolbox [11], which is based on [12]. It uses CART classification trees as weak learners. As before, we scale the input intensity values to the range [-1, 1], and the parameters are set using 10-fold cross-validation.

PCA+LDA. We applied PCA to the training set to obtain the eigenfaces [13, 14] with highest eigenvalues that explain at least 90% of the variance. In this subspace, LDA is applied to extract the projection that discriminates best between the two classes. The class of each test image is then predicted as the class with closest projected mean estimated in training [15].

2.2.2. Classification methods using invariant features

As an alternative to direct image classification methods, which require strong pre-processing, we use SIFT invariant features [5] extracted from the face images to train a second set of classifiers. The use of scale invariant features is standard in image classification, and it has been explored for gender classification in [3]. Due to the invariant properties of the SIFT keypoints, no histogram equalization, alignment or resize of the images is needed. So the only pre-processing performed was face localization using the face coordinates reported by [4]. We consider the following classifiers:

Bag Of Words: Visual dictionaries [2] is a general image classification technique consisting of an unsupervised and a supervised phase. In the unsupervised phase a clustering algorithm is run on the keypoints generally using a large number of clusters. The second phase trains a classifier on the histograms of visual word occurrences. In our experiments we use kmeans with 1000 clusters and SVM with a RBF kernel trained on the binary word histograms.

Evidence Random Trees. Evidence trees [6] is a two phase algorithm based on random forests. First, a random forest is built using the vector keypoints labelled with the class of the image they belong to. The trees store the training class histogram of the keypoints that belong to each terminal node. For each image the keypoints are dropped through all the trees of the ensemble and the terminal class histograms are accumulated and normalized. This histogram containing the evidence for each class is passed through a stacked classifier (bagging in this case) to compute the final class for the image.

Naïve-Bayes Nearest-Neighbor: Another effective technique based on Nearest Neighbor is Naïve-Bayes Nearest-Neighbor [7]. This method does not require a training phase, but only to keep all keypoints of all training images. A new image containing $d_1, d_2, ..., d_n$ keypoints is classified with the class k that minimises $\sum_{i=0}^{n} ||d_i - NN_k(d_i)||^2$, where NN_k computes the nearest keypoint in the training images that belong to class k.

Voted Nearest-Neighbor. Finally we consider a simple nearest neighbor classification approach. For each test image, with let us say n keypoints, we compute the k nearest neighbors of all these keypoints in the training set. The assigned class is then determined by majority voting over the $n \times k$ training keypoints, where each point votes for the class of the image it

Method	Error	# wins
PCA + LDA	$11.5\% \pm 2.9\%$	13
SVM	$10.0\% \pm 2.9\%$	42
ADABOOST	$12.1\% \pm 2.7\%$	15
NNet	$15.3\% \pm 4.3\%$	3
BOW	$19.8\% \pm 3.8\%$	0
NBNN	$10.7\% \pm 3.0\%$	38
ERT	$14.2\% \pm 3.1\%$	6
VNN	$13.1\% \pm 3.8\%$	6

Table 1. Average error and number of wins across the 100 random partitions.

Method	EER	AUC
PCA + LDA	$11.7\% \pm 3.0\%$	$94.7\% \pm 1.9\%$
SVM	$9.7\% \pm 2.8\%$	$96.4\% \pm 1.5\%$
ADABOOST g	$11.9\% \pm 2.7\%$	$95.2\% \pm 1.8\%$
NNet	$14.0\% \pm 4.1\%$	$92.3\% \pm 2.9\%$
BOW	$19.4\% \pm 4.0\%$	$89.2\% \pm 2.9\%$
NBNN	$9.7\% \pm 2.7\%$	$95.7\% \pm 1.8\%$
ERT	$15.0\% \pm 3.0\%$	$91.7\% \pm 2.7\%$
VNN	$11.3\% \pm 3.1\%$	$94.5\% \pm 2.2\%$

Table 2. Average EER and AUC across the 100 random partitions.

belongs to. The number k of neighbors is determined using a 10-fold cross-validation.

3. RESULTS

Tables 1 and 2 and Figure 1 show the performance of the different classification methods across the 100 partitions. In the second column of Table 1 we show the average classification error of each studied method. The third column shows the number of times each methods obtains the best accuracy in the 100 executions. When several methods obtain the best result in one execution, those methods are each assigned a "win". In Table 2, the average equal error rate (EER) and area under the curve (AUC) are shown. Figure 1 shows the average ROC for the best algorithms: SVM and NBNN.

The performance measures shown in these tables indicate that SVM and NBNN are the best methods followed by PCA+LDA and Adaboost. The performance of the other methods is somewhat inferior. It is important to note that there is not a clear trend with respect to what family of techniques is more adequate for gender classification. Methods based both on keypoints (NBNN) and on direct image classification (SVM) can obtain rather accurate results.

To compare the overall performance of the studied methods, we use the framework introduced in [16]. This framework permits to easily visualize the statistical differences

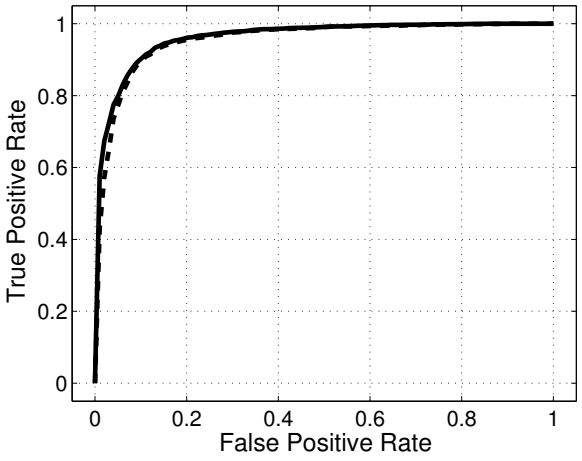

Fig. 1. Average ROC curves in test across the 100 random partitions of the best investigated methods. Solid: SVM. Dashed: NBNN.

among different algorithms. First, each method is ranked in each execution (rank 1 for the best method, 2 for the second and so on). Then a Nemenyi test is applied to compute the statistical differences among the methods. The results of this test are shown in Figure 2. The average rank obtained by each method is shown in lower axis. Methods for which the differences in average rank are not statistically significant with p-value < 0.05 are linked with a solid black line. Differences in average rank above the critical distance (CD) are considered significant. The CD is displayed at the top of the figure for reference.

From this figure it can be observed that SVM and NBNN are the best performing methods. The differences between these two methods and between NBNN and PCA+LDA are not statistically significant. A second group of algorithms includes ADABoost and VNN, which are close to the best methods but their performance is lower with statistical significance. The rest of methods have a poorer performance in the studied dataset.

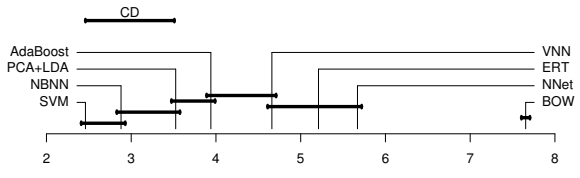

Fig. 2. Average ranks of each method

4. DISCUSSION

In this work, an evaluation of robust methods for gender classification is carried out. Two families of techniques are analysed: one based on image normalization and the other on invariant feature extraction (eight methods in total). Both approaches seek robust image classification against variations in illumination, scale orientation and so on. Additionally, in order to obtain reliable estimations of the different performance measures 100 random partitions of the data were used. A method that obtains the best performance in a single partition is not necessarily the best algorithm for the dataset. Carrying out a comparison in a single partition, which is a common practice, can produce different results (as shown from the number of wins of table 1) just by chance.

The best classification methods of the study are SVM trained on the normalized images and NBNN trained on SIFT keypoints. These two strategies provide statistically undistinguishable results. However, NBNN does not require image normalization, which has the advantages of reducing the level of image preprocessing, and avoiding the problems derived of the imperfections of current face alignment methods [4]. Furthermore, we expect that systems based on invariant features would be more robust to larger variations in scale, pose, orientation and illumination, as well as to partial occlusions, than systems based on image normalization strategies. This will be explicitly tested in our future work.

5. REFERENCES

[1] E. Mäkinen and R. Raisamo, "An experimental comparison of gender classification methods," *Pattern Recognition Letters*, vol. 29, pp. 1544–1556, 2008.

[2] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray., "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, ECCV 2004*, 2004, pp. 1–22.

[3] M. Toews and T. Arbel, "Detection, localization, and sex classification of faces from arbitrary viewpoints and under occlusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1567–1581, 2009.

[4] E. Mäkinen and R. Raisamo, "Evaluation of gender classification methods with automatically detected and aligned faces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 541–547, 2008.

[5] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int'l J. Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[6] G. Martínez-Muñoz, N. Larios, E. Mortensen, Wei Zhang, A. Yamamuro, R. Paasch, N. Payet, D. Lytle, L. Shapiro, S. Todorovic, A. Moldenke, and T.G. Dietterich, "Dictionary-free categorization of very similar objects via stacked evidence trees," in *IEEE Conferece on Computer Vision and Pattern Recognition, 2009. CVPR 2009*, 2009, pp. 549–556.

[7] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, 2008, pp. 1–8.

[8] 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.

[9] V. N. Vapnik, *The nature of statistical learning theory*, Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[10] C.C. Chang and C.J. Lin, "Libsvm: a library for support vector machines," 2010, `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

[11] A. Vezhnevets, "Gml adaboost matlab toolbox," 2010, `http://graphics.cs.msu.ru/ru/science/research/machinelearning/adaboosttoolbox`.

[12] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *The Annals of Statistics*, vol. 38, no. 2, pp. 337–374, 2000.

[13] L. Sirovich and M. Kirby, "Low -dimensional procedure for the characterization of human faces," *Journal of the Optical Society of America A*, vol. 4, pp. 519–524, 1987.

[14] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neuroscience*, vol. 3, pp. 71–86, 1991.

[15] S. Buchala, N. Davey, T.M. Gale, and R.J. Frank, "Principal component analysis of gender, ethnicity, age, and identity of face images," in *In IEEE ICMI*, 2005.

[16] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.