Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

An experimental comparison of gender classification methods

Erno Mäkinen *, Roope Raisamo

Multimodal Interaction Research Group, Tampere Unit for Computer–Human Interaction, Department of Computer Sciences, University of Tampere, FIN-33014, Finland

ARTICLE INFO

Article history:
Received 29 June 2007
Received in revised form 6 January 2008
Available online 6 April 2008

Communicated by W. Zhao

PACS:
4.001
5.002
5.009
5.010
5.016
5.017
7.001

Keywords:
Automatic face analysis
Gender classification
Face detection
Face database
Performance evaluation

ABSTRACT

Successful face analysis requires robust methods. It has been hard to compare the methods due to different experimental setups. We carried out a comparison study for the state-of-the-art gender classification methods to find out their actual reliability. The main contributions are comprehensive and comparable classification results for the gender classification methods combined with automatic real-time face detection and, in addition, with manual face normalization. We also experimented by combining gender classifier outputs arithmetically. This lead to increased classification accuracies. Furthermore, we contribute guidelines to carry out classification experiments, knowledge on the strengths and weaknesses of the gender classification methods, and two new variants of the known methods.

1. Introduction

The face is a challenging object to be recognized and analyzed automatically by a computer. However, successful analysis allows many interesting applications in human–computer interaction, security industry and psychology, among others. In addition, because many challenges associated with face recognition are common with other object recognition tasks, the solutions and knowledge in face recognition are also often applicable to other recognition problems.

Face analysis and recognition is receiving more and more attention. There are comprehensive surveys written on face recognition (Zhao et al., 2003) and on face detection (Hjelmås and Low, 2001; Yang et al., 2002). Facial expression analysis has also received attention and there are two extensive surveys on it (Fasel and Luettin, 2003; Pantic and Rothkrantz, 2000). Gender classification has been studied less. Gender classification has been especially interesting for psychologists but automatic gender classification has applications also in other fields, for example, in demographic data collection (Jain and Huang, 2004). Automatic gender classification is also a useful preprocessing step for face recognition since it is possible to halve, in case of equal amount of both genders, the number of face candidates before the recognition of the person and thus make the face recognition process almost twice as fast. In addition, separate face recognizers can be trained for the genders and in this way increase the face recognition accuracy. This has been successfully experimented in facial expression recognition by Saatci and Town (2006). Finally, the same methods can often be used both in gender classification and other face analysis tasks. For example, methods developed for gender classification can be applied to face recognition and vice versa.

The rising interest in face analysis also raises a need for reliable ways to compare and measure the methods. This allows researchers to put effort on the most promising methods and will advance the face analysis field more rapidly. Success in face analysis will enable applications in other fields, including entertainment and security, and will further increase interest towards face analysis.

We contribute to this field with a comprehensive comparison of the state-of-the-art gender classification methods. We

* Corresponding author. Tel.: +358 335518883; fax: +358 335516070.
E-mail address: etm@cs.uta.fi (E. Mäkinen).

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.03.016

experimented with several different methods and describe the results considering their classification accuracies and speed. The comparison is based on two image databases: the FERET database (Phillips et al., 1998, 2000) and an image database that we collected from the WWW. A problem with previous research on gender classification has been that the classification results have not been comparable because there has not been a common database to compare the methods, and the experiments have been carried out with different sets of images. In addition, the experimental setups have varied in many ways, making the comparison even harder. We present comparable results for gender classification. Finally, we propose guidelines to be followed when carrying out gender classification experiments and reporting the results. However, these guidelines are also applicable to other kind of face analysis experiments.

Gender classification is often used in connection with face detection, because this is needed in fully automatic face analysis systems. There are rather few studies where automatic face detection has been performed before gender classification. These are the reasons why we compared methods when automatic face detection is performed before the classification. For the face detection we used the cascaded face detector by Viola and Jones (2001). We chose it because it is both robust and a well known method for frontal face detection. However, we also experimented without automatic face detection. In this case we normalized face locations using manually located eyes before the gender classification to see how much classification accuracy is improved when compared to results gained with automatic face detection and to see if there are differences between the gender classification methods. In both cases we also experimented with how increasing the face image size by including hair in the images affects the classification accuracy. Finally, we experimented by arithmetically combining the gender classifier outputs. It could help, for example, with face images that the majority but not all of the classifiers classify correctly.

The article is organized in seven sections. After this introduction we describe the related work. We briefly discuss face detection, cover existing gender classification methods and describe face recognition research that is related to this paper. In the technical background section we introduce the cascaded face detector by Viola and Jones (2001) since to fully understand the gender classification results we present, it is mandatory to understand the overall functionality of the face detector. After describing the face detector we cover the compared gender classification methods. In the experimental setup section we introduce the face image databases used in the experiment and describe how we carried out the experiment. After reading the experimental setup section the reader can carry out experiments following the same test procedure as we did. The results of the experiment are described in the next section and after this, the results, including the differences between the methods, are discussed. We also propose a set of guidelines to be followed when carrying out similar experiments in the discussion section. Finally, we present some concluding remarks.

2. Related work

2.1. Face detection

Face detection methods can be roughly divided in to feature-based approaches and appearance-based approaches. In the feature-based approaches varying features are extracted from the image and face detection is based on the features. Such features can be, for example, edges or skin color and sometimes feature-based methods use knowledge on the face geometry (Mäkinen and Raisamo, 2002). In the appearance-based approaches, the whole image, possibly with some preprocessing, is used as an input to the face detector. Detailed analysis of different face detection methods can be read, for example, in the surveys by Hjelmås and Low (2001) and by Yang et al. (2002). In the next section we describe cascaded face detector by Viola and Jones (2001) in detail because it was used in our experiments.

2.2. Gender classification

Next, we will give an overview of the earlier research on gender classification and in the next section we give detailed description of the compared methods.

The research on automatic gender classification goes back to the beginning of the 1990s. The very first results were reported simultaneously by Cottrell and Metcalfe (1990) and by Golomb et al. (1990). In the same way as with face detection, gender classification methods can be roughly divided in feature-based and appearance-based methods. The first two methods were appearance-based and both used a multi-layer neural network approach. Faces were manually aligned for the experiments.

Many different methods have been tried after the two first methods were published. We describe them briefly by proceeding from the oldest research to the newest. Brunelli and Poggio (1995) experimented with HyperBF networks. They extracted a set of geometrical features from faces and used them as input to the networks that learned differences between the genders. The 79% classification rate for the novel faces was achieved without hair in the face images. Abdi et al. (1995) experimented with a radial-basis function (RBF) network and a perceptron with and without eigendecomposition as a preprocessing step. They found out that as good classification results can be achieved with pixel-based input as can be achieved with measurement-based inputs (such as the geometrical features used by Brunelli and Poggio (1995)).

Wiskott et al. (1997) presented a system where a model graph was placed manually to a face and gender classification was based on the Gabor wavelets placed on the model nodes. They also used the system for face recognition. Tamura et al. (1996) experimented with very low resolution face images and neural networks. They achieved 93% classification rate with only $8 * 8$ size face images. Lyons et al. (2000) used Gabor wavelets with principal component analysis (PCA) and linear discriminant analysis (LDA) to detect a face and classify gender. The graph similar to that of Wiskott et al. (1997) was placed on the face automatically. They achieved 92% classification rate that was a little bit higher than the 91.3% classification rate achieved by Wiskott et al. (1997).

Shakhnarovich et al. (2002) combined the cascaded face detector by Viola and Jones (2001) with threshold Adaboost (Freund and Schapire, 1997) trained classifiers for gender and ethnicity classification. There is a direct connection to our experiments because we used the cascaded face detector by Viola and Jones (2001) in the experiments and one of the compared gender classifiers is a threshold Adaboost classifier.

Sun et al. (2002) claimed that feature selection is an important issue for the gender classification and showed that genetic algorithms (GA) fit well for the task. They created feature vectors from face images using principal component analysis (PCA). Then they selected a subset of the features from the vectors using genetic algorithms and used the features as input to a gender classifier. Performance of four different classifiers was compared: Bayesian, neural network, support vector machine (SVM) and linear discriminant analysis (LDA). The SVM classifier achieved the best classification rate 95.3%.

Wu et al. (2003) used Look Up Table (LUT) based weak classifiers that were selected with Adaboost. Wu et al. (2003) showed that LUT Adaboost can model features that have multi-peak value distributions while the threshold Adaboost is constricted to single-peak distributions. We also experimented with LUT Adaboost classifiers to find out if LUT Adaboost is superior to threshold Adaboost in gender classification problem and to see how it performs compared to other kind of classifiers.

Jain and Huang (2004) used independent component analysis (ICA) to extract features from the face images and LDA to classify gender. They achieved impressive 99.3% classification rate with manually cropped and normalized FERET face images. Costen et al. (2004) used sparse SVM to classify genders. They achieved 94.42% classification rate with Japanese face images.

Sun et al. (2006) tried two different classifiers: Self Organizing Map (SOM) and threshold Adaboost. The novelty with their approach was that they used Local Binary Patterns (LBPs) to create features for the input. The best classification rate, 95.75%, was achieved with the Adaboost classifier. Also Lian and Lu (2006) experimented with LBPs. However, they used SVM as a classifier and achieved 96.75% classification rate. We decided to experiment somewhat similarly to Lian and Lu (2006) with SVMs and LBPs. However, we also experimented with pixel-based input in addition to LBP features.

Saatci and Town (2006) experimented with a SVM that was trained with the features extracted by an active appearance model (AAM). They had a two phase classifier. First the expression of the face was classified (categories: happy, sad, angry, neutral, and unrecognized). Then a gender classifier that was specific to the expression was used to recognize gender. This way they aimed to improve gender classification rate. However, the gender classification rate was decreased although they were able to improve facial expression classification rates by having separate expression classifiers for both genders. They suggested that the reason for the decrease in gender classification task was in the small amount of training images.

The methods described above have often achieved impressive classification performances and many of the methods are novel and interesting as such. However, comparisons of the methods have been hard because various and sometimes non-public databases have been used in the experiments. In addition, different normalizations such as face location normalization can be and have been used before classifying gender.

To our knowledge there are only two previous studies (Gutta et al., 1998; Moghaddam and Yang, 2002) where both public database and automatic face detection were used. Gutta et al. (1998) experimented with radial basis function (RBF) networks and inductive decision trees. However, they did not report the method used for the automatic face detection and face normalization. Moghaddam and Yang (2002) used maximum-likelihood estimation system for face detection that also did face alignment and contrast normalization automatically. For gender classification they used a support vector machine (SVM) with radial basis function (RBF) kernel. They used 1755 face images from the FERET database (Phillips et al., 1998, 2000). Moghaddam and Yang report 96.6% classification rate with good quality image data with their SVM system. However, according to Shakhnarovich et al. (2002) the classification rate of the SVM system for the data that was collected from the WWW was 75.5%. In addition, the cascaded detector architecture is about a 1000 times faster than the SVM system (Shakhnarovich et al., 2002).

Finally, Baluja and Rowley (2007) used pixel comparison operators with Adaboost classifier and achieved over 93% classification accuracy that surpassed SVM classifier accuracy that used pixels as input. The features they used are attractive choice for real-time systems that combine face detection and gender classification because the comparison operator values are fast to calculate. Baluja and Rowley (2007) achieved 50 times faster classifications with the Adaboost classifier than with the SVM. Baluja and Rowley (2007) also did sensitivity analysis for the classifiers by varying in-plane rotation, scale, and translation offsets of the face images. This kinds of inaccuracies are likely when faces are detected and, as shown by Baluja and Rowley (2007), also affect gender classification accuracies.

To gain insight in the differences between gender classification methods we carried out experiments with public FERET database face images (Phillips et al., 1998, 2000). We used both automatic face detection and manual face locating. We also studied how gender classification performance is affected when hair is included or excluded from the facial images. In addition, the classification performance of the methods were tested with a challenging WWW face database when automatic face detection was in use. Finally, we studied if an arithmetic combination of the gender classifier outputs would bring a performance gain.

2.3. Face recognition

In the face recognition field, there have been increasing efforts to attain comparable results for different methods. Probably the best known public database for face recognition is the FERET database (Phillips et al., 1998, 2000). It contains 14,051 face images of 1199 individuals. Depending on the person there are images with different facial expressions and poses. Also, there are images where illumination conditions have been changed or images are taken at different date. The FERET database has later been updated with the color version that contains 11,338 images of 994 individuals and is largely the same as the original FERET database but in color format.

There have also been several Face Recognition Vendor Tests (FRVTs) (Phillips et al., 2003, 2007) by US government that evaluate commercial and prototype face recognition technologies. The latest evaluation was FRVT 2006 (Phillips et al., 2007). As a part of the evaluation, US government provided the Biometric Experimentation Environment (BEE) that makes it easier for an experimenter to evaluate the methods and researchers to prepare their method for the evaluation.

The CSU Face Identification Evaluation System (CSU, 2003) provides implementations of some well known algorithms (PCA, PCA combined with LDA, Bayes, and Elastic Bunch Graph Matching with Gabor jets) and protocols to carry out face recognition experiments.

Although there have been efforts to ease comparison of the methods in the face recognition field, there have not yet been corresponding efforts in the gender classification. This is one of the main reasons why we did the comparison for many fundamentally different gender classification methods with different test setups.

3. Technical background

To understand the results presented in this article it is necessary to be familiar with the methods compared. Next we describe the relevant face detection and gender classification methods.

3.1. Cascaded face detector

The cascaded face detector (Viola and Jones, 2001) searches faces from an image by starting from the top left corner of the image and ending in the bottom right corner of the image (see Fig. 1). The image is searched through several times with a different sub-image size each time. When the image is scanned through, the sub-images are passed to the first layer of the face detector cascade to determine whether the sub-image contains a face or not (see Fig. 2) If the first layer classifies the sub-image as a face then the sub-image is passed to the next layer. This is continued until the sub-image is passed through all layers or discarded in a layer. If it is passed successfully through all layers then the final classification is a face and otherwise a non-face.

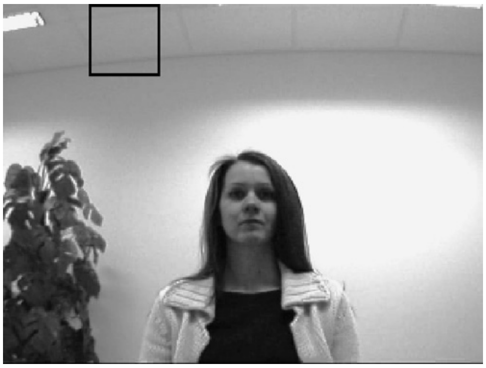

Fig. 1. Each image is scanned from top left corner to bottom right corner using sub-images.

The cascaded detector can be thought as a feature-based method, because it extracts simple rectangular features from the sub-images and makes detection decision based on the extracted features (see Figs. 2 and 3). The original cascaded detector used four types of rectangular features and each type is shown in Fig. 3. Later more feature types have been proposed (Lienhart and Maydt, 2002). Earlier layers have fewer features than the later layers. It also takes less time to process a sub-image in the earlier layers. The idea of the cascade is that promising regions where a face could occur are determined rapidly at the early layers of the cascade. During face detection, a value is calculated for a feature by subtracting sub-image pixel intensities of the dark rectangle from intensities in the white rectangle. Feature classifications are combined to one value in the layer. Some features in the layer have less effect than the others. The combined value is then compared to a predetermined threshold that determines if there is a face or not.

The detector by Viola and Jones (2001) also makes use of the Adaboost (Freund and Schapire, 1997) algorithm for the feature selection and integral image for the feature value calculation. The Adaboost will be described in the next section because it is used also with some of the compared gender classification methods. Further details of the cascaded detector can be found from the paper by Viola and Jones (2001). We have used the detector in this work because it can detect frontal faces in gray scale images reliably at the rate of 15 frames per second with a typical PC and it is also very well known.

3.2. Gender classification

Common to all gender classification methods is that each method takes the face sub-image as an input. However, the rest depends on the method and the details of the experiment. Next we describe each of the tested methods in such a detail that is needed to understand the experiments and results presented in this paper.

3.2.1. Neural network

Before the face image is inputted to a neural network there are some preprocessing phases. First the face image is scaled, if necessary, to a specific size. The size can be, for example, 48 $*$ 48 pixels. The number of input nodes in the network is equal to the amount of pixels in the resized face image. Then, the resized face image is histogram equalized, its intensity values are scaled to range from -0.5 to 0.5, and the values are stored in a vector. This vector is then given to the network that produces an output between -0.5 and 0.5. A negative value is defined as female classification and a positive value as male classification. See Fig. 4 for the illustration of how the output of the face detector is converted to the input of the neural network.

In addition to the input nodes and one output node, the network has a selected amount of hidden nodes. There are connections between the nodes and the connections have separate weights. An example of a multilayer neural network is shown in Fig. 5.

The neural network is trained by giving it female and male face example images as input. It is trained in rounds, so that all exam-

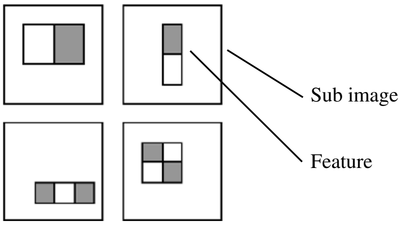

Fig. 3. Examples of rectangular features that are used in face detection.

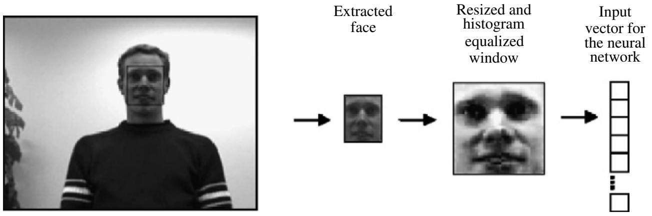
Extracted face
Resized and histogram equalized window
Input vector for the neural network

Fig. 4. The output of the face detector is converted to the input of the neural network.

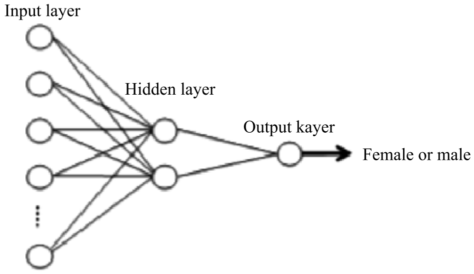

Fig. 5. A multilayer neural network.

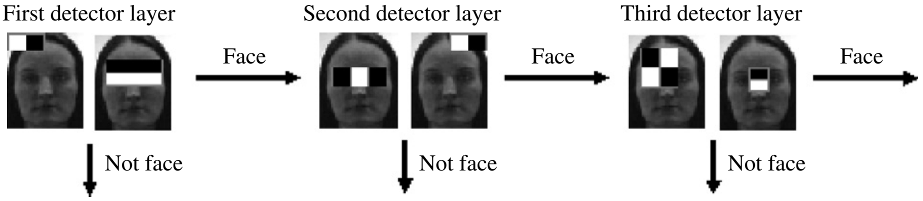

Fig. 2. A face detector cascade. For each layer two features are shown.

ples are inputted to it one by one in a round. The connection weights are changed after each image. When the image is inputted the next time to the network the output is closer to the expected output. The connection weights are changed using the back-propagation algorithm.

In addition to the training set of the example faces, a set of validation example faces are used. The validation images are inputted to the network after each training round and the output errors are calculated for the validation images. The error is calculated for an image by subtracting the output of the network for an image from the expected output and taking the absolute value of the result. For example, if the output of the network for a female example image is 0.2 then the error is 0.7 since the expected output is -0.5. The training is continued until the summed error for the validation images starts to increase.

In the experiments we tried different input layer sizes (images resized), hidden layer sizes and learning rates. The best combinations found are reported with the corresponding parameters in the experimental setup section.

3.2.2. Support vector machine (SVM)

The idea of the SVM is that samples (face images in this case) of different classes (female/male) can be separated in the higher dimensional space using the transformed features. Features that are transformed can be, for example, image pixel intensities or Gabor filter values (Lyons et al., 2000). For the transformation a kernel function is used. Many different kernel functions have been proposed in the literature but RBF and polynomial kernels are probably the most used ones.

We experimented with two fundamentally different SVM classifiers. In the first case, face image pixels were used as input to the SVM. We trained an SVM with histogram equalized image pixels the intensities of which were scaled to range from -1 to 1 and transformed with the RBF kernel. In the second case, local binary pattern (LBP) (Ojala et al., 1996) values were calculated from the images and used as an input to the SVM. An RBF kernel was also used in this case. We used LIBSVM (Chang and Lin, 2001) to train SVMs and predict classification rates for these two methods.

An SVM is trained so that such decision surface is searched that separates the training examples of separate classes with the maximum distance in the high dimensional space. There are several algorithms for the search and the search also depends of the selected kernel. We used the grid search provided with the LIBSVM (Chang and Lin, 2001). To avoid over-fitting a set of validation images can be used as with the neural network.

3.2.3. Local Binary Patterns (LBPs) combined with SVM

Local Binary Patterns (LBPs) (Ojala et al., 1996) are features that are calculated from pixel intensities in a pixel neighborhood. The basic idea is that as many binary values are created as there are pixels in the neighborhood of the center pixel. At the end these are concatenated to one binary value. Originally LBP was defined for $3 * 3$ pixel neighborhood but later it was extended to different neighborhoods and also some other modifications were done (Ojala et al., 2002).

Rotation invariant uniform patterns are an extension to the original LBP. They solve the practical problem that some patterns may occur too rarely to create reliable statistics for specific analysis problem. The formula and a more thorough description for basic LBP and rotation invariant uniform LBP calculation can be found from the paper by Ojala et al. (2002).

We decided to combine LBP with SVM in somewhat similar manner that was done by Hadid et al. (2004) for face recognition and by Lian and Lu (2006) for gender classification. We divided face image to $8 * 8$-blocks and filtered each block with the basic LBP operator with four neighbors at the radius of one ($LBP_{4,1}$). Then we created a histogram of each block. Since there are 16 different values that can be produced by the $LBP_{4,1}$, each histogram had 16 bins and each bin contained the amount of each value in the filtered block. We also filtered the whole face image with the uniform LBP with eight neighbors at the radius of one ($LBP^{u2}_{8,1}$) and created a 59-bin histogram for it. Finally, all the histograms were concatenated. For example, with a $24 * 24$ image there were nine $8 * 8$-blocks. So, total amount of bins in the concatenated histogram vector was $9 * 16 + 59 = 203$. The vector was used as an input to the SVM.

The difference between our variant of the method and the method by Lian and Lu (2006) is that they used $LBP^{u2}_{8,1}$-operator for face image that was divided in blocks and they did filtering only for the blocks and not for the whole image. We decided to do filtering also for the whole face image because Hadid et al. (2004) achieved better results in face recognition when the histogram of the whole face was used in addition to the blocks.

3.2.4. Discrete Adaboost

In the Adaboost algorithm (Freund and Schapire, 1997) specific features are selected. The weak classifiers that are used with the selected features form together a strong (reliable) classifier. The features and weak classifiers can be anything as long as they classify the given data examples (in this case face images) to specified classes (female/male). In the experiments we used haar-like features, three kinds of weak classifiers (threshold, mean and LUT), face images as data and two classes (female and male).

In the threshold weak classifiers, each weak classifier has a selected threshold value. When an image is classified with the weak classifier the value calculated for the image with the classifier is compared to the threshold. The classification is decided either as male or female depending on whether the calculated value is smaller or bigger than the threshold. The optimal threshold is selected during training so that the smallest possible number of example faces is misclassified with the feature.

We used simple rectangular features with the threshold weak-classifiers that are also used with the cascaded face detector. All the feature types used are shown in Fig. 3. The width and height of the features varied as well as locations in the face sub-image.

We experimented also with novel mean weak classifiers. They have a threshold that is calculated by first calculating mean feature values separately for the male and female training example faces. The threshold is then set to be halfway between the means. Otherwise the classification happens as with the threshold weak classifiers and we used the same set of rectangular features for both classifiers.

LUT Adaboost was developed by Wu et al. (2003). It differs from threshold and mean Adaboost (weak classifiers) so that instead of a threshold it places the calculated feature value to a bin. The number of bins is determined before the training, so that the range of the feature values is divided equally between the bins. The amounts of female and male examples that get a specific feature value are calculated for each bin during training. When the amount of the bins is increased more complicated feature value distributions can be modelled.

When a face image is classified the feature values are calculated for the face image and each value is compared to the bin corresponding to that feature and value. If the bin contains more female feature values (calculated during training) than male feature values then the classification result of that feature is female and otherwise male. The same set of rectangular features was used with the LUT Adaboost as with the threshold and the mean Adaboost.

So, there are six methods we experimented with: neural network, SVM, SVM with LBP features, threshold Adaboost, mean Adaboost and LUT Adaboost. The neural network and SVM are appearance-based methods while the other four are feature-based methods. The SVM with LBP has not previously been used in gender classification in the same way as we used. The mean Adaboost is a novel Adaboost variant.

4. Experimental setup

To experiment with the methods described in Section 3 we chose four different test setups. The setups as well as the databases used in the setups are described in this section.

4.1. Face databases

We used two image databases in the experiments: the FERET image database (Phillips et al., 1998, 2000) and a WWW image database. Examples from the FERET and WWW databases are shown in Fig. 6. The faces in Fig. 6 have been detected with the cascaded detector and histogram equalized.

The FERET database contains good quality gray scale images of 1199 individuals from different poses and with varying facial expressions. We used fa- and fb-subsets that contain frontal faces of 1196 people. Duplicate images of the same person were removed, so that only one image per person was left. We detected faces automatically from the images and removed false detections. Since the detected faces included more male faces than female faces, we randomly left some male faces out. The gender ground truths were added by a researcher. At the end we had 900 frontal face images from the database: 450 female and 450 male images.

The WWW image database contains 4720 frontal face images (2360 female and 2360 male images) that we randomly collected from the World Wide Web and annotated. The faces were detected automatically by the cascaded face detector and false detections were removed manually. The gender ground truths were added by a researcher.

Both databases have face images of people with varying ages, with eye glasses and without, with facial hair and without and of different ethnic backgrounds. However, the images in the FERET database are of good quality while images in the WWW database vary a lot in quality.

4.2. Test setups

We experimented with two appearance-based gender classification methods and four feature-based methods. The appearance-based methods were multi-layer neural network and SVM. Both methods take histogram equalized image pixels as an input. The feature-based methods were threshold Adaboost, LUT Adaboost (Wu et al., 2003), mean Adaboost, and a method that uses LBP values calculated from the face images as an input to a SVM. The mean Adaboost is a novel variant of the discrete Adaboost (Freund and Schapire, 1997) and the SVM with LBP has not been used in gender classification in the same way as we did.

We did comprehensive experiments with all the methods. First we tested the methods with face images that were detected by the face detector and scaled to the size of $24 * 24$ pixels. We used the cascaded detector and the default frontal face cascade provided with OpenCV (OpenCV, 2005) to detect faces. The minimum face size that the cascade detected was $24 * 24$ pixels and the sub-image size was scaled after each image scan by multiplying the current sub-image size with 1.25 (1st scan sub-image size: $24 * 24$ pixels, 2nd scan sub-image size: $30 * 30$ pixels, and so on).

Since some researchers (Abdi et al., 1995; Lyons et al., 2000) have achieved higher classification rates when face images with hair were used compared to those without hair, we decided to do that comparison also. We used the data described above but we increased the detected face area. We added 10% to the width on both sides (20% in total), 40% to the top and 12% to the bottom of the image since this provided us with face images that usually contained the hair in addition to the face but as little as possible any other data of the image, for example, the background. In some cases the area could not be grown as much as intended because image borders came across. In these cases we removed the image from the data used in the experiments. In some cases a hat or some other object, for example, a hand was in front of the hair or the image was otherwise considered to be of bad quality and was removed.

With the images that contained hair we used $32 * 40$ image size (instead of $24 * 24$ image size) because this size is closer to the image area growing percentages. After resizing the images we had 760 FERET images and 3808 WWW images and both image sets contained an equal amount of genders. Examples of the resized FERET and WWW database images with hair are shown in Fig. 7.

In addition, we experimented by normalizing the faces in the FERET images. By normalization we mean that eyes were located manually in the faces and faces were rotated and aligned in the images so that each face image had eyes in the same location and all the faces were in the upright position. The algorithm for the face normalization is shown in Table 1. The area around the eyes that was included in the face images was also determined with the algorithm.

After determining the face area with the normalization algorithm the areas were scaled to the size of $24 * 24$ or $32 * 40$ pixels, as with the automatically detected faces. With the $32 * 40$ size images we added 10% to width on the both sides (20% in total), 40% to the top and 12% to the bottom of the area before scaling, as was done with the automatic face detection. However, due to

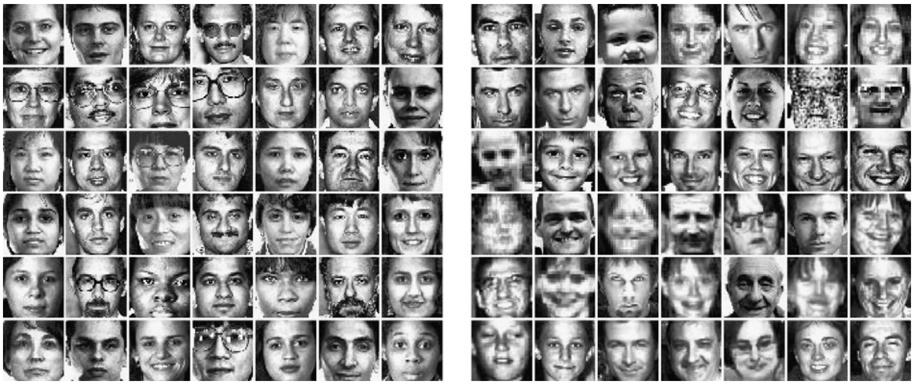

Fig. 6. Example face images from the FERET database on the left and from the WWW database on the right.

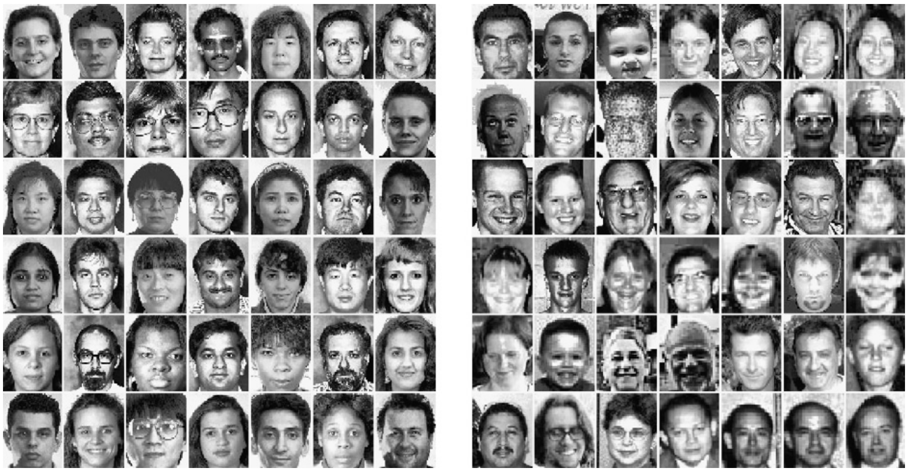

Fig. 7. Resized example face images including hair from the FERET database on the left and from the WWW database on the right.

Table 1
The algorithm for face normalization

(1) Eyes are located
(2) Image is rotated, so that eyes are vertically aligned
(3) Euclidean distance d_0 between the eyes is calculated in the rotated image
(4) Ratio r is calculated by $r = d_0/d_t$, where d_t is distance of the eyes in the resized image
(5) Width w_0 and height h_0 of the of the area around the eyes are calculated by $w_0 = r * w_t$ and $h_0 = r * h_t$, where w_t and h_t are width and height of the resized image (e.g. $w_t = h_t = 24$ or $w_t = 32$ and $h_t = 40$)
(6) Coordinates for the corners of the face area in the rotated image are calculated by $x_l = x_e - w_0/2$, $y_t = y_e - h_0/r_h$, $x_r = x_l + w_0$ and $y_b = y_t + h_0$, where x_l is x-coordinate of the left border, x_e is x-coordinate of the point in halfway between the eyes, y_t is y-coordinate of the top border, y_e is y-coordinate of the eyes, x_r is x-coordinate of the right border, y_b is y-coordinate of the bottom border and the ratio of the height above and below eyes r_h was 3.5

the rotation some face areas with the 32 * 40 size would have gone partially out of the original image bounds and we had to leave such faces out. This caused us to have 754 images for the test setup with the normalized images with hair.

At the end we had four test setups: images without hair and without normalization, images with hair and without normalization, images without hair and with normalization, and images with hair and with normalization. Table 2 summarizes the amount of images in each setup. The training was always done with the FERET images.

4.3. Experiments

First we experimented by using part of the FERET images in the training set. We put 80% of the FERET images in the training set. As test sets we used 20% of the FERET images. When doing the tests without image normalization we also put all the WWW images in the test set. Then we experimented by using 80% of the WWW images in the training set. All the FERET images and 20% of the WWW images were put to the test set.

Because with the neural network and SVM there is a danger of over-fitting to the training data, we used a part of the training images for validation when selecting optimal parameters for the methods. Since Adaboost is resistant to over-fitting (Freund and Schapire, 1997) we trained classifiers using Adaboost directly with the whole training set.

For the neural network we separated from 2% to 3% of the training images in the validation set and trained several neural networks using different amounts of hidden neurons and learning rates. We then tested each neural network with the test images. The best parameters found experimentally for different image sizes with and without hair are shown in Table 3 and in Table 4.

For the SVM we searched the best parameters with the fivefold cross-validation, so that 20% of the training images were in the validation set at a time. After the optimal parameters were selected we trained the final classifiers with the whole training set of the images. The best SVM parameters are shown in Table 3 and in Table 4. The best parameters are shown separately for the SVM with pixel inputs and for the SVM with the LBP features.

We selected 500 features for each Adaboost classifier. For the LUT Adaboost one can use different amount of bins. We experimented with 4, 6, 8 and 12 bins. Again, the best parameters, in this

Table 2
Number of the images in the test setups

	Not normalized		Normalized	
	Without hair (24 * 24)	With hair (32 * 40)	Without hair (24 * 24)	With hair (32 * 40)
Number of the FERET images	900	760	900	754
Number of the WWW images	4720	3808	–	–

Table 3
Best parameters for the methods when the FERET images were used for training and as test images

	Not normalized		Normalized	
	Without hair (24 * 24)	With hair (32 * 40)	Without hair (24 * 24)	With hair (32 * 40)
Neural network				
Number of hidden nodes	1	1	2	2
Input-hidden layer learning rate	0.04163	0.04163	0.0007211	0.0279399
Hidden-output layer learning rate	0.7071068	0.7071068	0.01	0.57735
SVM (RBF kernel)				
C	8.0	2.0	32.0	2.0
Γ	0.0078125	0.0078125	0.0001221	0.0078125
LBP + SVM (RBF kernel)				
C	8.0	2.0	2.0	2.0
Γ	0.03125	0.0078125	0.03125	0.03125
LUT Adaboost				
Number of bins	6	8	4	8

Table 4
Best parameters for the methods when the WWW images were used for training and as test images

	Not normalized	
	Without hair (24 * 24)	With hair (32 * 40)
Neural network		
Number of hidden nodes	1	1
Input-hidden layer learning rate	0.04163	0.0279399
Hidden-output layer learning rate	0.7071068	0.7071068
SVM (RBF kernel)		
C	8.0	8.0
Γ	0.0078125	0.001953125
LBP + SVM (RBF kernel)		
C	2.0	2.0
Γ	0.03125	0.03125
LUT Adaboost		
Number of bins	6	6

case the amount of bins, which produced best classification results, are shown in Table 3 and in Table 4.

After training, the classifiers were tested with the test set specific to the test setup. The results presented in the next section are based on the best performing parameters in each test setup.

5. Results

We measured classification rates of the methods in each test setup as well as how efficient they were when classification and training speed was considered. In this section we report the results of the experiments. The results are discussed in detail in the next section.

5.1. Classification rates with the FERET training images

The first presented results were gained when the FERET images were used for training. The classification accuracies with the FERET images are higher than with the WWW images. There are two probable reasons for this. The first one is that the FERET training set was more similar to the FERET test set than to the WWW test set. The second probable reason is that the WWW face images varied more in quality and were harder to classify.

The best classification rates for each method with FERET images without normalization are shown in Table 5 and with normalization in Table 6. The best classification rates for each method with WWW images are shown in Table 7.

The ROC curves for the FERET images in each test setup are shown in Fig. 8. The curve can be drawn for a method by changing the threshold value that determines the classification. For example, with the neural network we used the possible output values between −0.5 and 0.5. So, changing the threshold little by little from −0.5 to 0.5 we affect the fraction of faces classified as males and females. The closer the threshold is to the −0.5 the more female faces will be classified as female but at the same time more male faces will be classified as female. The fraction of males classified correctly is presented on the *y*-axis and the fraction of the females classified incorrectly is presented on the *x*-axis. For example, a curve point at the coordinates $(x,y) = (0.21, 0.84)$ means that 21% of the females are classified incorrectly when 84% of the males are classified correctly. The bigger the area under the ROC curve is the better the method is at classifying genders. The perfect curve would be such that it goes from the lower left corner to the upper left corner and from there to the upper right corner.

Table 5
Results for the FERET images without normalization when a separate set of the FERET images was used for training

Method	Classification rate %		
	Without hair (24 * 24)	With hair (32 * 40)	Average classification rate
Neural network	83.89	90.07	86.98
SVM	84.44	72.85	78.65
Threshold Adaboost	82.22	83.44	82.83
LUT Adaboost	80.56	87.42	83.99
Mean Adaboost	76.67	87.42	82.05
LBP + SVM	75.56	72.19	73.88
Average classification rate	80.56	82.23	81.40

Table 6
Results for the FERET images with normalization when a separate set of FERET images was used for training

Method	Classification rate %		
	Without hair (24 * 24)	With hair (32 * 40)	Average classification rate
Neural network	92.22	90.00	91.11
SVM	88.89	82.00	85.45
Threshold Adaboost	86.67	90.00	83.34
LUT Adaboost	88.89	93.33	91.11
Mean Adaboost	88.33	90.00	89.17
LBP + SVM	80.56	92.00	86.28
Average classification rate	87.59	89.56	88.57

Table 7
Results for the WWW images without and with hair when FERET images were used for training

Method	Classification rate %		
	Without hair (24 * 24)	With hair (32 * 40)	Average classification rate
Neural network	65.95	61.29	63.62
SVM	66.48	57.41	61.95
Threshold Adaboost	66.29	66.75	66.52
LUT Adaboost	66.19	64.81	65.50
Mean Adaboost	66.14	67.02	66.58
LBP + SVM	67.25	66.54	66.90
Average classification rate	66.38	63.97	65.18

5.2. Classification rates with the WWW training images

Since the classification rates were better for the FERET images than for the WWW images when the FERET images were used as training images one could consider what would happen if WWW images were used as training images. Would the WWW image classification accuracy increase or would the FERET image classification accuracy be decreased while nothing would happen to the classification accuracy of the WWW images or would there be some other effect? The results when using the WWW images as training images are shown in Tables 8 and 9.

The results presented in this subsection show that the WWW image classification accuracy increased about 10% units when over 3000 WWW images were used for the training. However, the FERET classification accuracies decreased from 5 to 10 percentage units when compared to corresponding results in the previous subsection. The results indicate that, although it is beneficial to have a large training image set, it is more important to have similar images in training and test sets than to have a large amount of training images.

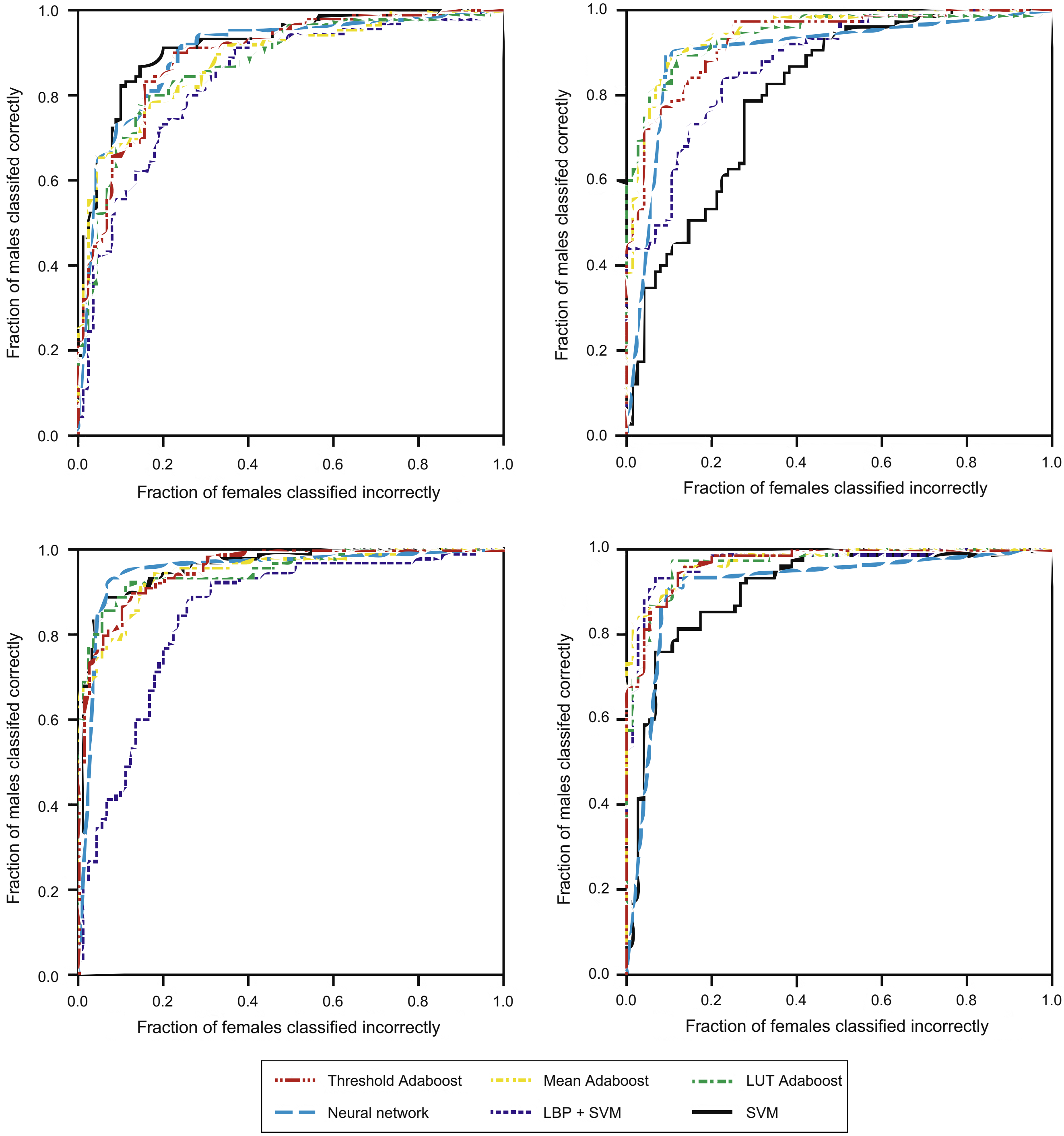

Fig. 8. ROC curves for the FERET images. (a) Images without hair (24 * 24 pixels), no normalization. (b) Images with hair (32 * 40 pixels), no normalization. (c) Images without hair (24 * 24 pixels), normalization. (d) Images with hair (32 * 40 pixels), normalization.

5.3. Combining gender classifier outputs

Individual gender classifier achieves classification accuracy that is far from perfect. One possibility to try to improve accuracy is to combine outputs of the several classifiers together. This can naturally be useful only if the gender classifiers make classification mistakes on different face images.

We combined all the classifier outputs together using four types of combination methods: voting, class voting, arithmetic, and arithmetic class voting. Voting was based on the amount of binary female and male classifications. Since there were six classifiers there could be equal amount of votes for females and males. In these cases the classification was not done. The number of such situations was small. For example, with the FERET test images with hair when FERET images were used for training 7.2% of the images were not classified. With the rest of the combination methods the classification was always done. The class voting did first vote between the SVM classifiers (pixel-based input and LBP features) and between the Adaboost classifiers, and then between neural network and the other two classes.

Arithmetic combination was done so that outputs of each classifier were normalized to the range between −0.5 and 0.5. For neu-

Table 8
Results for the FERET images without and with hair when WWW images were used for training

Method	Classification rate %		
	Without hair (24 * 24)	With hair (32 * 40)	Average classification rate
Neural network	79.00	67.50	73.25
SVM	79.44	75.00	77.22
Threshold Adaboost	74.00	72.76	73.38
LUT Adaboost	74.44	76.71	75.58
Mean Adaboost	74.44	74.08	74.26
LBP + SVM	73.44	69.21	71.33
Average classification rate	75.79	73.21	74.50

Table 9
Results for the WWW images without and with hair when a separate set of the WWW images was used for training

Method	Classification rate %		
	Without hair (24 * 24)	With hair (32 * 40)	Average classification rate
Neural network	73.62	60.26	66.94
SVM	78.28	75.13	76.71
Threshold Adaboost	75.32	75.26	75.29
LUT Adaboost	74.47	76.71	75.59
Mean Adaboost	70.13	71.84	70.98
LBP + SVM	75.32	76.71	76.01
Average classification rate	74.52	72.65	73.59

ral network output values were always in this range so nothing had to be done. The SVM classifiers produced probability between 0 and 1 and the output value had sign indicating the female or male classification, so the output value was divided by two. The Adaboost values could in theory be between 0 and 500 since there was 500 features but in practice they were found to be at most about 50 units distance from the threshold value. In average they were about 15 units from the threshold value. It was decided to divide the Adaboost output value by 100. The sign was determined by subtracting threshold value from the output value. After the outputs of each classifier were normalized they were summed together. If the sum was negative then the classification was determined to be female and if positive then male.

The arithmetic class voting happened so, that the arithmetic combination was used inside each class. The classes were same as with the class voting. After combination was done the voting happened between the classes.

The results for each combination are shown in Table 10. As can be seen all combinations produce better classification accuracies than any individual classifier. The voting method produces the best classification accuracy with all but one test image set and it has the highest average accuracy, so one could consider it as the best method if it is acceptable to have face images for which the classification is not done. If every face image should be classified then class voting would be the best choice based on these results as it has the second highest classification accuracy. However, it is good to note that if there would odd amount of classifiers then the voting method would classify all face images (but would not necessarily produce the best accuracies anymore).

5.4. Speed

The time taken in gender classification compared to the time that it takes to detect faces in an image is very small. In practice gender can be classified in real-time with all the methods when using a standard PC. However, there are huge differences in training times between the methods.

Table 10
Results for the classifier combination methods

Method	Classification rate %				
	WWW without hair	WWW with hair	FERET without hair	FERET with hair	Average
Best individual classifier	78.28	76.71	84.44	90.07	82.38
Voting	81.00	83.14	84.97	92.86	85.49
Class voting	79.55	81.59	85.71	90.78	84.41
Arithmetic	79.66	77.89	84.44	92.72	83.68
Arithmetic class voting	79.24	79.34	85.56	91.39	83.88

When FERET (or WWW) training set was used then also FERET (or WWW) test set was used. Images without normalization were used.

A neural network with 576 input nodes, 4 hidden nodes and 1 output node can be trained using back-propagation algorithm in a few minutes with a regular 1.79 GHz Pentium PC when there are hundreds or some thousands of the face images. However, the time consuming part with the neural network is to find optimal number of the neurons in each layer and the optimal learning rates, and the parameter selection has to be done manually. Also, several neural networks have to be trained even with the same set of parameters before the best network is found. So, training and selecting the best neural network can take many hours or even more time.

With SVM the kernel and search algorithm affect the training time. We used grid search in our experiments and it took about ten minutes to train a SVM with RBF kernel using 700 face images of size 24 * 24 pixels on a 1.79 GHz Pentium PC with 1 GB of memory. The amount of time was about the same with LBP features although the LBP-features had a little bit less data.

It took about half an hour to train an Adaboost classifier with 500 features with a parallelized algorithm using a 4 processor IBM pSeries 690 computer and 700 face images. Using a regular PC with Pentium III processor the training would take a day or a few days.

The training time may be an issue, for example, when selecting a method for a commercial product or even when doing academic research. The Adaboost training is the most time consuming one but as it was pointed out also neural network and SVM training can take time. However, there are no great differences in the classification speeds between the methods especially when the longer time required for face detection is considered. If using classifier combinations to achieve higher gender classification accuracy this still holds.

6. Discussion

6.1. Results

The most interesting thing in the results is that there were surprisingly small differences in the classification rates between the methods (Figs. 8 and 9). There were no statistically significant differences between the methods ($F_{5,18} = 1.22$, not significant) when calculated from the four test setups with the FERET images. This can also be seen from the error bars in Fig. 9 since they clearly overlap.

LBP features used with SVM produced the best average results with the WWW images (see Table 7), although with the FERET images that had not been normalized and with the FERET images that had been normalized but did not include hair it produced the worst classification rate. Also, SVM with image pixels as an in-

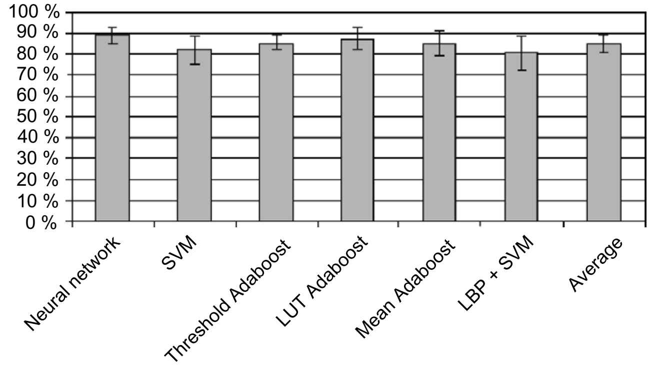

Fig. 9. Average classification rates for the methods with the FERET test images when a separate set of FERET images was used for training. The rates are averages for the methods in the all four test setups.

put had the best classification rate when the FERET images without hair and normalization were used but with the FERET images with hair it had the worst classification rate. It had the worst classification rate whether the normalization was used or not.

In some previous studies (Shakhnarovich et al., 2002; Sun et al., 2002; Wu et al., 2003) the differences between some of the experimented methods have also been relatively small. As our results show the differences between the methods are so small that the differences have no statistical significance and the methods performed statistically equally well.

The second quite surprising issue, considering the earlier research (Abdi et al., 1995), is that larger face areas with hair do not necessarily produce better classification results. However, we note that also Abdi et al. (1995) reported that, for example, face shape affects the classification and not only hair. The average classification rates for the methods with the FERET images in the four test setups are shown in Fig. 10. The classification rate was a little bit higher in average when a larger face area was used but not much and the difference was not statistically significant ($t_{22} = 0.843$, not significant).

However, since the classification rates of over 90% have been reported earlier in the cases where face location normalization was used (Gutta et al., 1998; Jain and Huang, 2004; Lyons et al., 2000; Moghaddam and Yang, 2002; Saatci and Town, 2006; Sun et al., 2002, 2006; Wiskott et al., 1997; Wu et al., 2003) it seems that this preprocessing is necessary to take a full advantage of the larger face area including hair and to attain classification rates of over 90%. Our results support this conclusion since there was a statistically significant difference in rates between the test setups with normalization and the test setups without normalization ($t_{22} = 3.555$, $p < 0.005$).

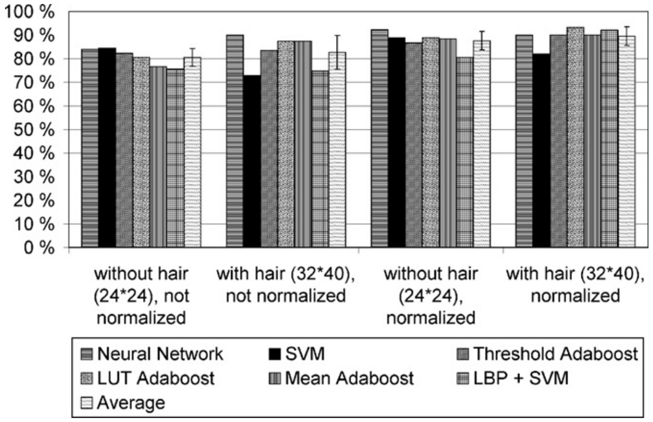

Fig. 10. Average classification rates for the methods with the FERET test images when a separate set of the FERET images was used for training. The rates are averages for the methods in the all four test setups.

Looking at the results it also becomes clear that the location normalization based on the eyes is more effective than just including the hair in the face images. The average classification rate for the methods when the FERET images without hair and with location normalization were used was 87.59% while for the FERET images with hair that have not been normalized it was 82.67%, although the difference in the rates was not statistically significant ($t_{10} = 1.475$, not significant).

In addition, although there were no statistically significant differences in rates between the images with and without hair that have not been normalized ($t_{10} = 0.640$, not significant), the average rate was higher for the images with hair. This was also true between normalized images with and without hair ($t_{10} = 0.867$, not significant). Anyhow, the best results were gained when both face location normalization and face images with hair were used. The average classification rate in this case was 89.56% and the difference for the images without hair and without normalization (the average recognition rate was 80.56%) was statistically significant ($t_{10} = 4.066$, $p < 0.005$). All these results support the fact that the face location normalization is more important than just including hair in the face images.

What also becomes clear is that our WWW database that contained images collected randomly from the WWW was rather challenging even when compared to the earlier research that utilized WWW image databases (Sun et al., 2006; Wu et al., 2003). It is clear that more preprocessing, including location and rotation normalization, would be needed to attain good recognition rate with the WWW images. The performance became even worse for the WWW images with a larger area that includes hair, possibly because backgrounds varied a lot in the WWW images and more background was present in the face images with hair (while in the FERET images there was only little variation in the backgrounds.) However, classification accuracy with the WWW images was increased when a set of WWW images was used for classifier training. The increase probably happened because the training images resembled more the test set than the high-quality FERET training images did. The other probable reason for increased accuracy is the amount of training images. Nevertheless, since the classification accuracy for the FERET images was decreased when the WWW images were used as training images it seems that the image set similarity is more important factor for the accuracy than the amount of training images. This should be taken in to account when, for example, creating a face analysis system with automatic face detection and gender classification that is to be used in a real application.

It proved to be useful to combine classifier outputs to gain increased classification accuracy. Using more than one gender classifier in classification decreases the classification speed but when classification is preceded by face detection the more important factor for the speed is face detection speed. Because face detection is often inaccurate the detected face can be rotated, scaled, and translated from the optimal when it is inputted to the classifier. As Baluja and Rowley (2007) have shown these inaccuracies can have drastic effect to the gender classification accuracy. One way to increase accuracy is to make classification based on the consensus decision by the classifiers.

6.2. Guidelines

Based on our analysis of the results, we suggest a set of guidelines to carry out other similar experiments. We feel that there is a need for this because there is increasing amount of research on the face analysis field and based on our experience sometimes more effort could have been put to achieve comparable results. In addition, these guidelines can also be applied to other classification problems, for example, to facial expression classification. Suggested guidelines to carry out the experiments are the following:

(1) Select at least one publicly available image database for the experiments. This ensures that other researchers can compare their results with your results more easily.
(2) Select databases that contain a large set of images. This ensures that classification rates are not produced by chance. This also guarantees that statistical tests can be performed to compare alternative recognition methods.
(3) Select at least two databases if you want to ensure that your method works in varying conditions. The other should contain a set of good quality images and the other a set of varying quality images. FERET (Phillips et al., 1998, 2000) is an example of a good quality image database and the WWW database we used is an example of a varying quality database. If your method is supposed to be used only with good quality images, then, of course, the good quality database can be used alone. More than two databases may be needed if you want to test the methods in a specific situation.
(4) Define precisely how you carry out the experiments. In other words, what are the parameters you use with the methods, do you use cross-validation, and if yes, which images are in the training and test sets in each round, and so on.
(5) Report precisely how you carried out the experiments and which parameters you used with the methods. This will help others to evaluate your results and put them in the right context.
(6) Be prepared to provide further information to fellow researchers. Inevitably, no matter how precise you are when reporting the results, there will be something unclear or forgotten. You might even be asked to provide implementation of your method. It is up to you if you want to provide it but, in any case, it is to everyone's advantage that the methods and results are comparable.

7. Conclusions

We carried out experiments on several state-of-the-art gender classification methods including two new variants. We compared the methods when automatic face detection was used as a preprocessing step and separately when manual alignment of the faces was done. We used two databases, namely the FERET database (Phillips et al., 1998, 2000) and a WWW image database to compare the methods. We found that when hair was included in the face images, it did not guarantee a better classification rate when compared to the face images without hair but face location normalization was more efficient and guaranteed an improved recognition rate. Second, we found that the used image database affected a lot the classification rates achieved and this should be taken into account when carrying out experiments. Third, the classification accuracies were increased when outputs of the classifiers were arithmetically combined. The experiments with the two databases brought new insight into gender classification with various methods in various conditions.

Acknowledgements

This work was funded by Tampere Graduate School in Information Science and Engineering. Portions of the research in this paper use the FERET database of facial images collected under the FERET program.

References

Abdi, H., Valentin, D., Edelman, B., O'Toole, A.J., 1995. More about the difference between men and women: Evidence from linear neural network and principal component approach. Neural Comput. 7 (6), 1160–1164.

Baluja, S., Rowley, H.A., 2007. Boosting sex identification performance. Internat. J. Comput. Vision 71 (1), 111–119.

Brunelli, R., Poggio, T., 1995. HyberBF networks for gender classification. In: Proc. DARPA Image Understanding Workshop, pp. 311–314.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: A library for support vector machines.

Costen, N., Brown, M., Akamatsu, S., 2004. Sparse models for gender classification. In: Proc. Internat. Conf. on Automatic Face and Gesture Recognition (FGR'04), May, pp. 201–206.

Cottrell, G.W., Metcalfe, J., 1990. EMPATH: Face, emotion, and gender recognition using holons. In: Lippmann, R., Moody, J.E., Touretzky, D.S. (Eds.), Proc. Advances in Neural Information Processing Systems 3 (NIPS). Morgan Kaufmann, pp. 564–571.

CSU, 2003. The CSU face identification evaluation system.

Fasel, B., Luettin, J., 2003. Automatic facial expression analysis: A survey. Pattern Recognition 36 (1), 259–275.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55 (1), 119–139.

Golomb, B.A., Lawrence, D.T., Sejnowski, T.J., 1990. SEXNET: A neural network identifies sex from human faces. In: Lippmann, R., Moody, J.E., Touretzky, D.S. (Eds.), Proc. Advances in Neural Information Processing Systems 3 (NIPS). Morgan Kaufmann, pp. 572–579.

Gutta, S., Wechsler, H., Phillips, P.J., April 1998. Gender and ethnic classification of face images. In: Proc. Internat. Conf. on Automatic Face and Gesture Recognition (FGR'98), pp. 194–199.

Hadid, A., Pietikäinen, M., Ahonen, T., 2004. A discriminative feature space for detecting and recognizing faces. In: Proc. Internat. Conf. on Computer Vision and Pattern Recognition (CVPR'04), vol. 2, pp. 797–804.

Hjelmås, E., Low, B.K., 2001. Face detection: A survey. Computer Vision and Image Understanding 83, 236–274.

Jain, A., Huang, J., May 2004. Integrating independent components and linear discriminant analysis for gender classification. In: Proc. Internat. Conf. on Automatic Face and Gesture Recognition (FGR'04), pp. 159–163.

Lian, H.-C., Lu, B.-L., 2006. Multi-view gender classification using local binary patterns and support vector machines. In: Proc. 3rd Internat. Sympos. on Neural Networks (ISNN'06), Chengdu, China, vol. 2, pp. 202–209.

Lienhart, R., Maydt, J., September 2002. An extended set of haar-like features for rapid object detection. In: Proc. Internat. Conf. on Image Processing (ICIP'02), vol. 1, pp. 900–903.

Lyons, M., Budynek, J., Plante, A., Akamatsu, S., 2000. Classifying facial attributes using a 2-d Gabor wavelet representation and discriminant analysis. In: Proc. Internat. Conf. on Automatic Face and Gesture Recognition (FG'00), IEEE, Grenoble, France, pp. 202–207.

Mäkinen, E., Raisamo, R., 2002. Real-time face detection for kiosk interfaces. In: Proc. APCHI 2002, Beijing, China, pp. 528–539.

Moghaddam, B., Yang, M.-H., 2002. Learning gender with support faces. IEEE Trans. Pattern Anal. Machine Intell. 24 (5), 707–711.

Ojala, T., Pietikäinen, M., Harwood, D., 1996. A comparative study of texture measures with classification based on featured distributions. Pattern Recognition 29 (1), 51–59.

Ojala, T., Pietikäinen, M., Mäenpää, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Machine Intell. 24 (7), 971–987.

OpenCV, 2005. OpenCV beta 5, Open Source Computer Vision Library.

Pantic, M., Rothkrantz, L.J., 2000. Automatic analysis of facial expressions: The state of the art. IEEE Trans. Pattern Anal. Machine Intell. 22, 1424–1445.

Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.J., 1998. The FERET database and evaluation procedure for face recognition algorithms. Image Vision Comput. J. 16 (5), 295–306.

Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J., 2000. The FERET evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Anal. Machine Intell. 22 (10), 1090–1104.

Phillips, P., Grother, P., Micheals, R., Blackburn, D., Tabassi, E., Bone, J., 2003. FRVT 2002 evaluation report. Tech. Rep. NISTIR 6965, National Institute of Standards and Technology.

Phillips, P.J., Scruggs, W.T., O'Toole, A.J., Flynn, P.J., Bowyer, K.W., Schott, C.L., Sharpe, M., 2007. FRVT 2006 and ICE 2006 large-scale results. Tech. Rep. NISTIR 7408, National Institute of Standards and Technology, Gaithersburg, March, MD 20899.

Saatci, Y., Town, C., April 2006. Cascaded classification of gender and facial expression using active appearance models. In: Proc. 7th Internat. Conf. on Automatic Face and Gesture Recognition (FGR'06), pp. 393–400.

Shakhnarovich, G., Viola, P.A., Moghaddam, B., 2002. A unified learning framework for real time face detection and classification. In: Proc. Internat. Conf. on Automatic Face and Gesture Recognition (FGR'02). IEEE, pp. 14–21.

Sun, Z., Bebis, G., Yuan, X., Louis, S.J., December 2002. Genetic feature subset selection for gender classification: A comparison study. In: Proc. IEEE Workshop on Applications of Computer Vision (WACV'02), pp. 165–170.

Sun, N., Zheng, W., Sun, C., Zou, C., Zhao, L., 2006. Gender classification based on boosting local binary pattern. In: Proc. 3rd Internat. Symposium on Neural Networks (ISNN'06), Chengdu, China, vol. 2, pp. 194–201.

Tamura, S., Kawai, H., Mitsumoto, H., 1996. Male/female identification from 8 to 6 very low resolution face images by neural network. Pattern Recognition 29 (2), 331–335.

Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR'01), vol. 1, pp. 511–518.

Wiskott, L., Fellous, J.-M., Krüger, N., von der Malsburg, C., 1997. Face recognition by elastic bunch graph matching. In: Sommer, G., Daniilidis, K., Pauli, J. (Eds.), 7th International Conference on Computer Analysis of Images and Patterns, CAIP'97, Kiel. Springer-Verlag, Heidelberg, pp. 456–463.

Wu, B., Ai, H., Huang, C., June 2003. LUT-based Adaboost for gender classification. In: Proc. Internat. Conf. on Audio and Video-based Biometric Person Authentication (AVBPA'03), Guildford, United Kingdom, pp. 104–110.

Yang, M.-H., Kriegman, D., Ahuja, N., 2002. Detecting faces in images: A survey. IEEE Trans. Pattern Anal. Machine Intell. 24 (1), 34–58.

Zhao, W., Chellappa, R., Phillips, P., Rosenfeld, A., 2003. Face recognition: A literature survey. ACM Comput. Surveys (CSUR) 35 (4), 399–458.