Evaluation of Gender Classification Methods with Automatically Detected and Aligned Faces

Erno Mäkinen and Roope Raisamo

Abstract—We present a systematic study on gender classification with automatically detected and aligned faces. We experimented with 120 combinations of automatic face detection, face alignment, and gender classification. One of the findings was that the automatic face alignment methods did not increase the gender classification rates. However, manual alignment increased classification rates a little, which suggests that automatic alignment would be useful when the alignment methods are further improved. We also found that the gender classification methods performed almost equally well with different input image sizes. In any case, the best classification rate was achieved with a support vector machine. A neural network and Adaboost achieved almost as good classification rates as the support vector machine and could be used in applications where classification speed is considered more important than the maximum classification accuracy.

Index Terms—Classifier design and evaluation, computer vision, face and gesture classification, face detection, interactive systems, machine learning, vision I/O.

1 Introduction

In principle, it is fairly easy to combine face detection and gender classification methods. The face is detected with the detector and then input into the gender classifier that determines the gender. However, this process is more complex than it appears and includes many aspects for consideration. The most important factors are usually the detection and classification accuracies. The other important factors are detection and classification speeds.

The issues to be considered are the selection of the detector and classifier, the features that are input of the detected face to the gender classifier, which (if any) normalization is used before gender classification, and if there is some processing that can be made common for both detection and classification. We examined the connection between face detection and gender classification experimentally.

In the next section, we describe the related work. Then, the combination of the detector and classifier is considered from the technical perspective. Next, the experiments and their results are described and the results of the experiments are discussed. Finally, we present some concluding remarks.

2 Related Work

There exist only a few studies where gender classification has been combined with automatic face detection. These studies have proposed novel methods for gender classification. We contribute by evaluating combinations of gender classifiers and alignment methods where faces have been automatically detected.

To the best of our knowledge, Moghaddam and Yang [1] developed the first automatic system for combined face detection and gender classification. They used maximum-likelihood estimation for face detection and for facial feature detection. For gender classification, they used several different classifiers. The experiments were carried out with a set of FERET images [2]. The most interesting findings in the context of this paper were that the Support Vector Machine (SVM) performed better than the other classifiers and resolution of the face did not much affect the classification rate with the SVM.

Shakhnarovich et al. [3] combined the cascaded face detector by Viola and Jones [4] with discrete Adaboost-based [5] gender and ethnicity classification. The advantage of the system by Shakhnarovich et al. [3] is that many preprocessing and bookkeeping calculations can be shared by the face detector and the gender classifier. This resulted in about 1,000-fold speed up when compared to the SVM classification [3]. Shakhnarovich et al. [3] collected images from WWW for the experiments.

Castrillón-Santana et al. [6] combined their real-time face detector ENCARA [7] with gender and identity classifiers. In the experiments they used Web camera images. The other work by Castrillón-Santana et al. [8] concentrated on a classifier that evolved while it was processing video image. They found that this kind of classifier was not as reliable as their corresponding offline trained SVM classifier and suggested several possible reasons for this. One reason was that, with the evolving classifier, eyes were automatically located while, with the offline classifier, they were manually located. Face alignment based on the eyes is one issue that we have analyzed carefully to find out the importance of face alignment when automatic face detection and gender classification are combined.

Wu et al. [9] used a cascaded detector with LUT Adaboost to detect faces. The detector was based on the cascaded detector by Viola and Jones [4]. After face detection, eye centers and mouth center were detected using Simple Direct Appearance Model (SDAM) [10], [11] and the face was aligned using the eyes and the mouth. After normalization, gender was classified with the LUT Adaboost classifier. Wu et al. [9] used a combination of FERET [2] images and images collected from WWW in the experiments.

BenAbdelkader and Griffin [12] presented a method that extracted regions from the face and used them as input for an SVM or Fisher Linear Discriminant (FLD) gender classifier. Faces were aligned using the eye locations. For the experiments, they used images obtained from various databases.

Yang et al. [13] reported a detailed analysis of how different normalizations affect gender classification accuracy. They had three different methods for alignment and three gender classifiers: an SVM, an FLD, and a two-layer Real Adaboost classifier. They used Chinese face images in the experiments. The most interesting fact from the viewpoint of this paper is their claim that shape-free alignment may produce better classification results with methods that use local features such as haar-like features, while shape-preserving alignment methods may produce better results with global features.

Finally, Baluja and Rowley [14] presented an Adaboost system for gender classification with manually aligned faces. They carried out a thorough experimental comparison between the Adaboost and an SVM classifier by varying face image scaling, translation, and rotation. We conducted a similar analysis and have compared their results to ours where appropriate.

3 Technical Background

The face detection method, the automatic alignment methods, and the gender classification methods used in the experiments are described next. They are described in such detail that the experiments can be understood.

3.1 Face Detection

For face detection we used the detector provided with OpenCV 1.0 [15] that is based on the cascaded detector by Viola and Jones [4]. It is one of the best detectors currently available in terms of speed and reliability and also well known. The detection takes place in such a way that the image is scanned through with image windows to find faces. The detector has a cascade of classifier layers. Each layer

- *The authors are with the Department of Computer Sciences, University of Tampere, FIN-33014, Finland. E-mail: {etm, rr}@cs.uta.fi.*

Manuscript received 5 Mar. 2007; revised 6 July 2007; accepted 25 Sept. 2007; published online 9 Oct. 2007.
Recommended for acceptance by M.-H. Yang.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2007-03-0142.
Digital Object Identifier no. 10.1109/TPAMI.2007.70800.

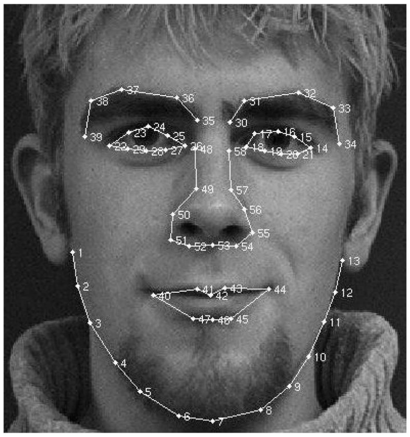

Fig. 1. An example face with defined landmarks and edges. The example face is from the IMM database [18].

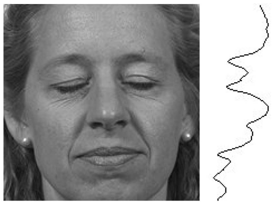

Fig. 2. An example of the detected face with the vertical intensity profile. The example face is from the FERET database [2].

contains an Adaboost classifier that takes haar-like feature values as input. An image window is passed through the cascade to determine if the window contains a face or not. OpenCV provides implementation of the cascaded detector with various trained classifier cascades. We used the default frontal face cascade in our experiments.

3.2 Facial Feature Detection and Face Normalization

We chose to experiment with four automatic alignment methods: three based on Active Appearance Model (AAM) [16] and one based on profile alignment [17]. In addition, we used histogram equalization with neural network and SVMs. Variance normalization was used for the integral face images used with the Adaboost method.

3.2.1 Active Appearance Model (AAM) Based Face Alignment

The Active Appearance Model (AAM) [16] is a statistical model of the shape and texture of an object, in this case of the face. We used AAM-API version 120506 by Stegmann et al. [18] both for model building and shape fitting. The model is built so that landmarks and edges are defined for a set of example faces. Fig. 1 shows an example. The model learns the general face shape and its variations by minimizing the model shape distance to the example faces.

The model fitting is done in such a way that the AAM shape is superimposed on the face (e.g., on the middle of the face) and landmarks of the shape are iteratively moved toward positions where they are (hopefully) closer to the real landmarks. The alignment is done with the shape landmarks.

3.2.2 Profile Alignment

The profile alignment is based on the eye locations found. The algorithm for locating the eyes is described in detail in our earlier article [17].

Eye detection using our method is as follows:

1. Vertical intensity profile of the found face is created (see Fig. 2).
2. Horizontal intensity profiles are created on the rows containing local minimum or maximum on the vertical profile and the horizontal profiles are smoothed using the adjacent rows.
3. Local minima are searched for from the horizontal profiles created on the local minimum rows to detect eyes. Nose and mouth are searched for in a similar way.
4. Based on the fuzzy rules and general knowledge of facial anatomy, for example, that the eyes are above the nose and the nose is above the mouth in an upright face, the best eye, nose, and mouth candidates are selected.

TABLE 1
Learning Rates Used with the Neural Networks

Layer	Number of nodes	Learning rate
Hidden-output layer	1 hidden node	0.7071068
Hidden-output layer	2 hidden nodes	0.577350
Hidden-output layer	10 hidden nodes	0.301511
Hidden-output layer	20 hidden nodes	0.2182179
Input-hidden layer	$24 * 24 = 576$ input nodes	0.0416305
Input-hidden layer	$36 * 36 = 1296$ input nodes	0.0277671
Input-hidden layer	$48 * 48 = 2304$ input nodes	0.0208288

After facial features have been detected, the face alignment is done based on the eyes. The eyes are always detected in the same image row and this means that faces are not rotated but scaled and placed so that the located eyes are in the same horizontal positions in all face images. Also, eye candidates are not always found and, in that case, alignment cannot be done. The advantage of the profile alignment in any case is that it is very fast compared to AAM alignment.

3.3 Gender Classification

We selected four gender classification methods: a multilayer neural network with pixel-based input, an SVM with pixel-based input, a discrete Adaboost with haar-like features, and an SVM with LBP features [19].

3.3.1 Multilayer Neural Network with Image Pixels as Input

The multilayer neural networks that we used took histogram equalized face pixels as input. The input values were scaled to range from -0.5 to 0.5. There were as many input nodes as there were pixels in the face images. For example, with the image size $24 * 24$ pixels, there were 576 input nodes. We used one hidden layer with one, two, 10, or 20 hidden nodes and output layer with one node. The output was between -0.5 and 0.5. The output value above zero was defined as male and the value below as female.

The neural network was trained using the standard backpropagation algorithm. As learning rates, we used the values that we found to be good and they are shown in Table 1.

3.3.2 Support Vector Machine (SVM) with Image Pixels as Input

The second classifier was an SVM [20] that took histogram equalized face pixel intensity values scaled to range from -1 to 1 as input. The RBF kernel was used in the transformation. We used the SVM implementation provided with LIBSVM version 2.82 [21] in the experiments.

3.3.3 SVM with LBP Features as Input

Local binary patterns (LBPs) were introduced and later extended by Ojala et al. [19], [22]. The basic idea of LBPs [19], [22] is that binary values are calculated from a pixel neighborhood and the binary values are concatenated to one binary value.

We decided to combine LBP with SVM in a manner somewhat similar to that of Lian and Lu [23]. We divided the face image into blocks. With $24 * 24$ images, there were nine $8 * 8$-blocks and, with

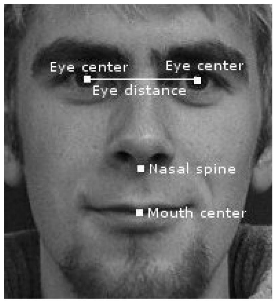

Fig. 3. Decision on successful alignment is based on the facial landmarks and the eye distance. The example face is from the IMM database [18].

$48 * 48$ images, there were 36 $8 * 8$ blocks. With face images of size $36 * 36$, it is not possible to divide images evenly into $8 * 8$ blocks. In this case, we used $4 * 8$ blocks at the right side of the images, $8 * 4$ blocks at the bottom of the images, and one $4 * 4$ block at the bottom right corner of the images.

Each block was filtered with the basic LBP operator with four neighbors at a radius of one ($LBP_{4,1}$). We then created a histogram for each block. Since 16 different values 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 a radius of one ($LBP^{u2}_{8,1}$) and created a 59-bin histogram for it. The total number of bins in the concatenated histogram vector was $n * 16 + 59$, where n was the number of blocks in the image. The vector was used as an input to an SVM.

3.3.4 Discrete Adaboost with Haar-Like Features

Discrete Adaboost [5] performs classification based on a set of weak classifiers and features. We used threshold weak classifiers and haar-like features, so that one haar-like feature corresponded one weak classifier. The haar-like feature types used were the same as those used by Viola and Jones [24].

4 Experiments

We studied the effects of face alignment, face image sizes, and various controlled misalignments on gender classification accuracy.

4.1 Data

We used the IMM Face database [18] and the FERET database [2] for the experiments. The AAM model was built of 14 face images, seven female faces and seven male faces from the IMM database. We used the Jacobian training scheme and truncated the shape, texture, and combined models so that they explained 95 percent of the variance present in the 14 faces.

We also tried building the AAM model from 50 FERET faces that we had annotated manually for facial landmarks but it emerged that the alignment results were clearly worse for the unseen FERET face images with this model. This is probably due to the fact that the FERET images that we used for building the model varied somewhat more in poses and lighting conditions than the IMM database images. We have, in any case, made the FERET image annotations available at our Web site http://www.cs.uta.fi/hci/mmig/vision/datasets/ for those who are interested.

Although the AAM model was built from the IMM database faces, the FERET image database was used when we tested the actual face detection, face alignment, and gender classification combinations.

We used 304 FERET face images, an equal number of both genders, for training the gender classifiers. These images were such that AAM alignment and profile alignment were successfully applied to the detected faces. The determination of the successful alignment was such that:

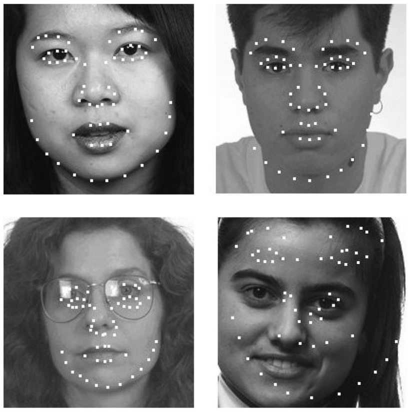

Fig. 4. Successful alignments in the top row and unsuccessful alignments in the bottom row. The example faces are from the FERET database [2].

1. The real euclidean distance between left and right eye center was calculated.
2. The euclidean distances from the automatically determined facial feature locations to the real ones were calculated.
3. The ratios of the distances to real eye distance were calculated with the equation $r = d_f/d_e$, where r was the calculated ratio, d_f was the feature distance, and d_e was the real eye distance.
4. The calculated ratios were compared to predetermined ratios and, if all the calculated ratios were smaller than the predetermined ratios, then the automatic alignment was deemed successful and otherwise unsuccessful.

The features we used for determining the successful alignment were the eye centers, the nasal spine, and the mouth center for AAM alignment, and the eye centers for profile alignment. If the eye centers were not located with the profile alignment, the face location determined directly by the face detector was used. In Fig. 3, we show the facial landmarks and the eye distance used for determining successful alignment. The predetermined ratio was 0.25 for the eye centers,and 0.2 for the nasal spine and the mouth center. In Fig. 4, we show some successful and unsuccessful alignments based on the above rules and ratios.

To test the trained gender classifiers, we used 107 FERET images (60 male and 47 female faces). Naturally, the badly aligning face images were not removed from these images since we wanted to find out if the alignment is useful when used in combination with face detection and gender classification.

For the out-of-plane rotation sensitivity test, we used FERET images of 56 males and 56 females as test images. There were nine poses per person (from $-60°$ to $+60°$).

The names of the images used during training and testing as well as the locations of the faces detected, facial feature locations defined, and genders are available at our Web site: http://www.cs.uta.fi/hci/mmig/vision/datasets/.

4.2 Procedure

We had several test variables that were varied in the experiments and thereby producing the 120 combinations each with unique test conditions. The variables and their conditions are shown in Table 2.

At first, we trained the gender classifiers using automatically detected and, in some combinations, aligned faces. If an alignment method was used then it was used both during training and testing. Only successfully detected and aligned faces were used for training. We trained separate classifiers for each combination.

For each combination having the neural network, we trained networks with one hidden neuron, with two hidden neurons, with 10 hidden neurons, and with 20 hidden neurons. The networks with one and two neurons were as reliable as the larger ones but

TABLE 2
Test Variables and Their Conditions

Variable	Conditions
Gender classification method	SVM with LBP features Neural network with face pixels SVM with face pixels Adaboost with haar-like features
Alignment method	None Manual Profile AAM with eyes AAM with eyes and nasal spine AAM with eyes and mouth
Input image size[1]	24*24 36*36 48*48
Timing of alignment[2]	Alignment before resizing face image Alignment after resizing face image

worked faster. In the reported results, we have used the rates of the better performing network with one or two neurons.

When AAM shape was fitted to a face, the maximum number of iterations allowed was 50. The initial position of the AAM shape was horizontally in the middle of the detected face and vertically slightly below the detected face center because a part of the jaw tended to be below the detected face area. The initial shape was also scaled so that the shape had the same width as the detected face. The eye centers were calculated from shape points 14, 18, 22, and 26 and the mouth center from points 42 and 46 (see Fig. 1.)

The faces were rotated and aligned so that eyes were at the same locations in all faces. If nose or mouth location was used then the face was stretched or squeezed so that the nose or mouth center was at the same vertical location in all aligned faces. For manual alignment, we used only the eye locations. The face area was determined after alignment with the following equations:

$$\begin{aligned} w_f &= h_f = d_e/10 * 24, \\ l &= (x_l + x_r)/2 - w_f/2, \\ r &= l + w_f, \\ t &= y_e - h_f/3.5, \\ b &= t + h_f, \end{aligned}$$

where w_f was the width of the face area, h_f was the height of the face area, d_e was the euclidean distance between the eyes, l, r, t, and b defined the left, right, top, and bottom borders for the face box, and x_l, x_r, and y_e were the coordinates for the eyes.

We trained the Adaboost classifiers with the whole training set. All of the Adaboost classifiers had 500 features. With the neural networks, we separated 2 percent of the training images to the validation set and trained them for 1,000 rounds. The neural network of the round with the lowest validation error was selected. For the SVM, the best parameters were searched using gird search and five-fold cross-validation. Twenty percent of the training images were in the validation set at the time. After the best parameters were found, the final SVM classifier was trained with the whole training set.

We used successfully detected faces as test images. Alignment was not required to be successful, but such face images were removed from the test set that would have led face area to be partially or wholly out of the image bounds after alignment. The out-of-plane rotation test set was separate. For this, automatic face detection was not used and gray backround (intensity value 127) was added when necessary after calculating the face area.

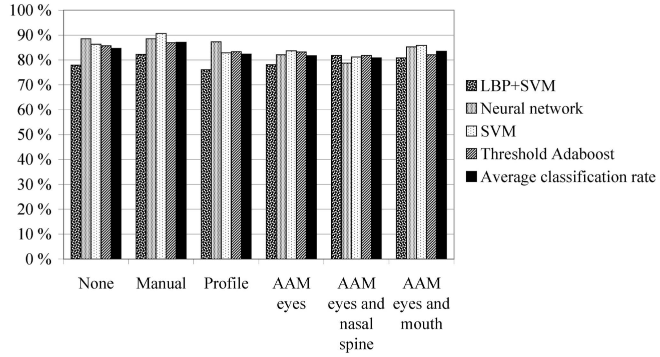

Fig. 5. Classification rates for the methods with different alignments.

4.3 Results

The average gender classification rate for all cases where automatic alignment was used was 82.1 percent. A possibly somewhat surprising result was that the average gender classification rate was higher when alignment was not used. It was 84.6 percent with no alignment, as shown in Fig. 5. The difference was also statistically significant (Wilcoxon signed-rank test: $z_{12} = 2.118, p = 0.034$). If we consider AAM and profile alignment separately, there was no statistically significant difference for AAM alignment ($z_{12} = 1.804, p = 0.071$) but, for profile alignment, there was ($z_{12} = 2.002$, $p = 0.045$). Anyhow, the average classification rate for AAM alignment was 82.0 percent and for profile alignment 82.4 percent, which were both smaller when compared to "no alignment" cases. The only classification method that had some increase in the classification performance when automatic alignment was in use was SVM with LBP features and even then only when AAM alignments were used. Manual alignment gave slightly better classification results (87.1 percent) than "no alignment" on average, but the difference was not statistically significant.

When comparing the automatic alignments to each other, the AAM alignment with eyes and mouth produced the best average classification rate. The AAM alignment with eyes and nasal spine resulted in the poorest classification rate. However, the differences were not large and the difference of the AAM eyes and mouth alignment to the AAM eyes and nasal spine alignment was the only statistically significant one ($z_{24} = 2.583, p = 0.010$.)

All of the methods achieved the best classification rate with manually aligned faces, as can be seen from Fig. 5. However, with the neural network, an equal classification rate was achieved without alignment. The methods had the poorest classification rate with the AAM eyes and nasal spine alignment with the exception of the SVM with LBP features, which had the poorest performance with the profile alignment. When the methods were compared, the SVM with pixel-based input, the neural network, and the Adaboost revealed statistically significant differences to the SVM with LBP features when using automatic alignment ($z_{24} = 3.018$, $p = 0.003$, $z_{24} = 2.415$, $p = 0.016$, and $z_{24} = 2.420$, $p = 0.016$ respectively.)

As can be seen from Table 3, the alignment methods worked best if they were used before resizing face image and the result was statistically significant ($z_{48} = 3.668, p = 0.0002$). As can be seen from Table 4, the situation was the same for all gender classification methods.

The classification rates for different face image sizes are shown in Fig. 6 and Table 5. The best classification rates were achieved with the image size $36 * 36$, but the difference from images of size $48 * 48$ was not statistically significant. However, when the image size $36 * 36$ was compared to the image size $24 * 24$, the difference was statistically significant ($z_{40} = 3.263, p = 0.001$), likewise when images of size $48 * 48$ were compared to images of size $24 * 24$ ($z_{40} = 1.960, p = 0.050$).

Differences in classification rates for gender classification methods were fairly consistent with different image sizes with the exception of the SVM with LBP features. For the SVM with LBP

TABLE 3
Classification Rates When Using Different Alignments and Alignment Is Done Before or After Resizing the Face

	Classification rate %		
Alignment	**Alignment before resizing**	**Alignment after resizing**	**Average**
Profile	83.80	80.92	82.36
AAM eyes	82.32	81.15	81.74
AAM eyes and nasal spine	82.40	79.28	80.84
AAM eyes and mouth	84.74	82.24	83.49
Average	83.32	80.90	82.11

TABLE 4
Classification Rates for Gender Classification Methods When Alignment Is Done Before or After Resizing the Face

	Classification rate %		
Method	**Alignment before resizing**	**Alignment after resizing**	**Average**
LBP+SVM	81.46	76.87	79.17
Neural network	84.66	81.93	83.30
SVM	84.35	82.40	83.38
Threshold Adaboost	82.79	82.40	82.60
Average	83.32	80.90	82.11

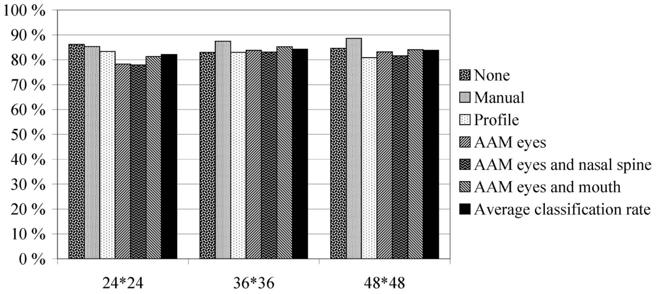

Fig. 6. Classification rates with diffrent alignments and face sizes.

TABLE 5
Classification Rates with Different Classifiers and Face Sizes

	Classification rate %			
Method	**24*24**	**36*36**	**48*48**	**Average**
LBP+SVM	76.92	79.07	82.06	79.35
Neural network	84.21	85.89	82.90	84.33
SVM	82.62	86.54	84.02	84.39
Threshold Adaboost	81.50	84.58	83.93	83.34
Average	81.31	84.02	83.23	82.85

TABLE 6
Alignment Accuracy Measurements

Alignment	**Image size**	d_{eye}	Δ_s	$\lvert\Delta_x\rvert$	$\lvert\Delta_y\rvert$	$\lvert\Delta_\alpha\rvert$
No alignment	N/A	N/A	N/A	0.036	N/A	2.451
Profile	N/A	0.196	1.201	0.056	0.060	2.451
	24*24	0.280	1.081	0.060	0.226	2.154
	36*36	0.278	1.062	0.045	0.223	2.445
	48*48	0.231	1.096	0.050	0.155	2.641
AAM with eyes	N/A	0.109	0.964	0.040	0.054	1.799
	24*24	0.149	0.957	0.080	0.037	2.353
	36*36	0.135	0.966	0.072	0.037	2.767
	48*48	0.128	0.963	0.067	0.041	2.400
AAM eyes and nasal spine	N/A	0.461	0.964	0.041	0.428	1.799
	24*24	0.560	0.957	0.079	0.502	2.353
	36*36	0.722	0.966	0.071	0.675	2.766
	48*48	0.706	0.963	0.068	0.656	2.400
AAM eyes and mouth	N/A	0.368	0.964	0.036	0.340	1.799
	24*24	0.439	0.957	0.072	0.386	2.353
	36*36	0.461	0.966	0.068	0.408	2.766
	48*48	0.474	0.963	0.064	0.424	2.400

features the classification rate increased when image size increased. For the remaining methods, the best classification rate was achieved with an image size of $36 * 36$ pixels. One should note, however, that, with the neural network, the second best classification rate was achieved with the $24 * 24$ pixel size images.

From the above one could assume that automatic alignment actually decreases the accuracy of the face location. To examine this issue, we calculated accuracy measurements based on the real and estimated eye locations. We used the relative eye location measure, d_{eye}, by Jesorsky et al. [25] and the measures by Rodriguez et al. [26]. The measures by Rodriguez et al. [26] were Δ_x and Δ_y for the horizontal and vertical alignment errors of the eyes, Δ_s for the scaling error, and Δ_α for the rotation error. The measures were calculated with the following equations:

$$\begin{aligned}
d_{eye} &= max(d(C_l, \tilde{C}_l), d(C_r, \tilde{C}_r))/d(C_l, C_r),\\
\Delta_x &= \overline{\mathrm{d}x}/d(C_l, C_r),\\
\Delta_y &= \overline{\mathrm{d}y}/d(C_l, C_r),\\
\Delta_s &= d(\tilde{C}_l, \tilde{C}_r)/d(C_l, C_r),\\
\Delta_\alpha &= \widehat{\overrightarrow{C_lC_r}, \overrightarrow{\tilde{C}_l\tilde{C}_r}},
\end{aligned}$$

where $d(a, b)$ is the euclidean distance between locations a and b, C_l and C_r are the real locations of the left eye and right eye, and $\tilde{C}_l$ and $\tilde{C}_r$ are the estimated eye locations. The $\overline{\mathrm{d}x}$ is horizontal distance from the center between the estimated eyes to the center of the real eyes and the $\overline{\mathrm{d}y}$ is the corresponding vertical distance. The Δ_α measure is the angle between the line intersecting the estimated eye centers and the line intersecting the real eye centers. Here, we have reported the averages of the absolute values of Δ_x, Δ_y, and Δ_α because the most interesting issue is how much the values differ from zero on average.

The measurements are shown in Table 6. For the Δ_s, the value 1 is best, and for the other measurements, the value 0. Only the Δ_x and Δ_α measurements are shown for the "no alignment" condition because only the horizontal center between the eyes and the angle of the detected face can be accurately calculated. The horizontal center of the eyes could be assumed to be horizontally at the center of the face box returned by the detector. The angle of the detected face was assumed to be 0° because the detector returned only upright face boxes.

The Δ_x seems to affect the classification accuracy more than the Δ_α, at least with the magnitude of errors shown here. The sensitivity analysis in the following section supports this reasoning. When the "no alignment" condition is compared to the other alignment conditions only the "AAM eyes with mouth" condition together with "alignment before resizing" condition had equal $|\Delta_x|$ and smaller Δ_α. The d_{eye} and $|\Delta_y|$ were smaller for all conditions where alignment occured before resizing when compared to the conditions where alignment occured after resizing except the $|\Delta_y|$ with the "AAM alignment with eyes" condition. Finally, "AAM eyes and mouth" alignment was more accurate than "AAM eyes and nasal spine" alignment. All of these differences in alignment measurements are also visible in the classification rates.

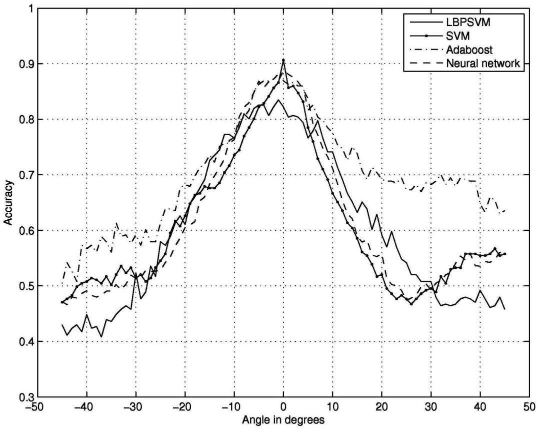

Fig. 7. Sensitivity of the classifiers to in-plane rotation.

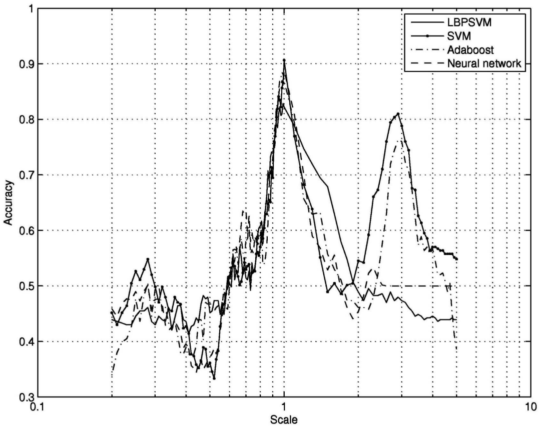

Fig. 9. Sensitivity of the classifiers to scaling.

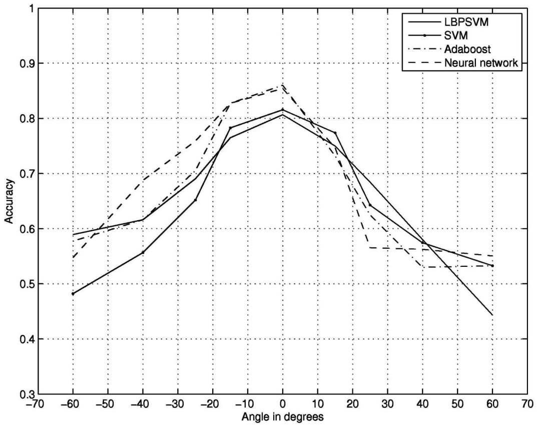

Fig. 8. Sensitivity of the classifiers to out-of-plane rotation.

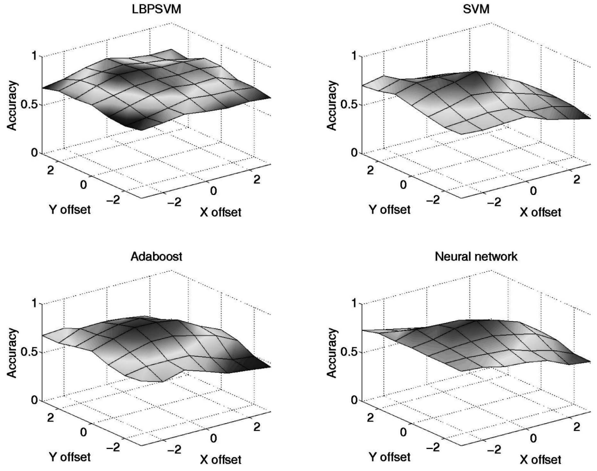

Fig. 10. Sensitivity of the classifiers to translation.

4.4 Sensitivity Analysis

The final issue studied was the sensitivity of the gender classifiers to various detection and alignment inaccuracies as did Baluja and Rowley [14]. We used manually aligned faces and varied the in-plane rotation of the faces from -45° to 45°, out-of-plane rotation from -60° to 60°, the scale from 0.2 to 5, and the translation in the horizontal and vertical directions from -3 to $+3$ pixels. The results of the rotation, scaling, and translation are shown in Fig. 7, Fig. 8, Fig. 9, and Fig. 10. The results are averages over all three image sizes.

The Adaboost classifier with haar-like features was the most resistant classifier feature combination for in-plane rotation variations. Out-of-plane rotation decreased classification accuracies more slowly than in-plane rotation. However, classification accuracy curves between the methods were fairly similar for out-of-plane rotation. For the scale, there are clear peaks in accuracies with the SVM and Adaboost near scaling factor 3, and below the chance classification accuracies with all classifiers between scaling factors 0.3 and 0.6.

5 Discussion

As the results show, the automatic alignment did not increase the gender classification rates. The effect was the same with all automatic alignment methods whether the alignment was done before or after image resizing. Also, with the exception of the SVM with LBP features, the gender classification method did not affect the results. If we take into account that manual alignment actually did increase the classification rates, the problem seems to be that the automatic alignment methods were just not reliable enough. One possible way to improve alignment would be to tune the parameters for the alignment methods. Adding more faces to AAM model building could help. One possibility would be to use some other alignment method than what we used. Wu et al. [9] used the Simple Direct Appearance Model (SDAM) [10], [11] for alignment. Yet another possible way to improve the alignment results could be to use several methods for locating facial features and do the alignment based on the results of all methods. The results also support the statement by Shakhnarovich et al. [3] that face alignment "is not entirely robust in the presence of significant pose, lighting, and quality variation" since the FERET [2] images had some variation in face poses and imaging conditions (although, for example, WWW images would have more).

If we consider the effect of face image resizing on alignment and gender classification, the alignment should be done before resizing. The image size after resizing was not that important. The best classification rates were achieved with an image size of $36 * 36$ pixels, but the difference in classification rates to image size of $48 * 48$ pixels was not statistically significant. However, Wu et al. [9] also achieved the best gender classification rate with images of size $36 * 36$ pixels. In addition, our results support the statement by Moghaddam and Yang [1] that the performance of SVM depends mainly on the number of training images and not so much on the input resolution. This also seems to hold for the other gender classification methods. The SVM with LBP features had an interesting characteristic that the classification rate increased when face image size was increased.

Sensitivity analysis revealed that Adaboost with haar-like features was the most resistant method to the in-plane rotation variations. In the results reported by Baluja and Rowley [14], the Adaboost classifier with the pixel comparison features was more resistant to in-plane rotation than the SVM with pixel-based input. It is likely that the features used are a more important factor than the classifier for the in-plane rotation sensitivity. This reasoning is based on the fact that the SVM with pixel-based input and the neural network with pixel-based input had very similar in-plane rotation sensitivity curves while the SVM with the LBP features had a different curve. An interesting fact is that the classification accuracies decreased faster for in-plane rotation than for out-of-plane rotation. When the scale was varied, the accuracy peak for the Adaboost was visible in our results and in the results by Baluja and Rowley [14] although the peaks appeared with different scales and we also had an accuracy peak for the SVM with pixel-based input. The below the chance accuracies that appeared with certain scales were also visible in the both results although again partly at different scale intervals. All the classifiers behave simlarly when the translation was varied.

The speed of the combined system is also an important issue. The face detection takes most of the time especially with large images, so the speed of the detector is the most important factor. If alignment is used, then the profile alignment is faster than the AAM alignment. However, rotated faces may become a problem with the profile alignment unless it is modified to be able to handle rotated faces and which would decrease the speed of the alignment. Of the gender classification methods, the Adaboost with haar-like features is slightly faster than the neural network, but both methods are much faster than SVM. The Adaboost method can be combined effectively with the cascaded detector, especially if no alignment is used [1].

What is therefore the best combination for automatic face detection and gender classification? The best classification rates were achieved with the SVM using pixel-based input when images of size $36 * 36$ pixels were used and when alignment was not used. However, the neural network performed equally well in practice and Adaboost came close, too. In any case, if the classification rate is the most important issue, then we recommend SVM, because the best classification rates have been reported for it in many studies [1], [6], [12], [13]. However, Adaboost achieves the highest classification speed and the neural network offers very good compromise between speed and accuracy. Further, there are indications in our results and in the results by Baluja and Rowley [14] that haar-like features and pixel comparison features are resistant to in-plane face rotation variations. This would make these features an attractive choice for many real applications. If automatic alignment is used, the alignment should be performed before resizing the face image.

6 Conclusion

We carried out an experimental evaluation on gender classification. The study included comparison of four fundamentally different gender classification methods and four automatic alignment methods together with nonaligned faces and manually aligned faces. We also analyzed how the classification accuracy was affected when face image resizing occured before or after alignment. Finally, we conducted a sensitivity analysis for the classifiers by varying rotation, scale, and translation of the face images.

We found that the SVM with $36 * 36$ pixel face images as input achieved the best gender classification rate. Although the automatic alignment methods implemented were not accurate enough to facilitate gender classification, the best classification rates were achieved with manually aligned faces and, by improving the implementation of the automatic alignment methods, one would achieve better classification rates. There are several ways one could try to improve automatic alignment. For example, more than two facial feature locating methods could be used and the alignment could be based on the consensus decision by the methods.

Acknowledgments

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

[1] B. Moghaddam and M.-H. Yang, "Gender Classification with Support Vector Machines," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 306-311, Mar. 2000.

[2] 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 J.,* vol. 16, no. 5, pp. 295-306, 1998.

[3] G. Shakhnarovich, P.A. Viola, and B. Moghaddam, "A Unified Learning Framework for Real Time Face Detection and Classification," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 14-21, 2002.

[4] P. Viola and M.J. Jones, "Robust Real-Time Face Detection," *Int'l J. Computer Vision,* vol. 57, no. 2, pp. 137-154, 2004.

[5] Y. Freund and R.E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J. Computer Systems Science,* vol. 55, no. 1, pp. 119-139, 1997.

[6] M. Castrillón-Santana, O. Déniz-Suárez, J. Hernández-Sosa, and A. Domínguez-Brito, "Identity and Gender Recognition Using the ENCARA Real-Time Face Detector," *Proc. Conf. Assoc. Espaola para la Inteligencia Artificial,* 2003.

[7] M.F. Castrillón-Santana, "On Real-Time Face Detection in Video Streams: An Opportunistic Approach," PhD dissertation, Universidad de Las Palmas de Gran Canaria, Mar. 2003.

[8] M. Castrillón-Santana, O. Déniz-Suárez, J. Lorenzo-Navarro, and M. Hernández-Tejera, "Gender and Identity Classification for a Naive and Evolving System," *Proc. Second Workshop Multimodal User Authentication,* 2006.

[9] B. Wu, H. Ai, and C. Huang, "Real-Time Gender Classification," *Proc. Third Int'l Symp. Multispectral Image Processing and Pattern Recognition,* vol. 5286, pp. 498-503, Oct. 2003.

[10] X. Xiao, "Face Detection and Retrieval," master's thesis, Tsinghua Univ., 2002.

[11] T. Wang, H. Ai, and G. Huang, "A Two-Stage Approach to Automatic Face Alignment," *Proc. Third Int'l Symp. Multispectral Image Processing and Pattern Recognition,* H. Lu and T. Zhang, eds., pp. 558-563, 2003.

[12] C. BenAbdelkader and P. Griffin, "A Local Region-Based Approach to Gender Classification from Face Images," *Proc. 2005 IEEE CS Conf. Computer Vision and Pattern Recognition,* p. 52, 2005.

[13] Z. Yang, M. Li, and H. Ai, "An Experimental Study on Automatic Face Gender Classification," *Proc. 18th IEEE Int'l Conf. Pattern Recognition,* vol. 3, pp. 1099-1102, Aug. 2006.

[14] S. Baluja and H.A. Rowley, "Boosting Sex Identification Performance," *Int'l J. Computer Vision,* vol. 71, no. 1, pp. 111-119, 2007.

[15] "OpenCV 1.0, Open Source Computer Vision Library," http://www.intel.com/technology/computing/opencv/, 2006.

[16] T. Cootes and C. Taylor, *Statistical Models of Appearance for Medical Image Analysis and Computer Vision,* 2001.

[17] E. Mäkinen and R. Raisamo, "Real-Time Face Detection for Kiosk Interfaces," *Proc. Asia-Pacific Conf. Computer-Human Interaction 2002,* pp. 528-539, 2002.

[18] M.B. Stegmann, B.K. Ersbøll, and R. Larsen, "FAME—A Flexible Appearance Modelling Environment," *IEEE Trans. Medical Imaging,* vol. 22, no. 10, pp. 1319-1331, 2003.

[19] T. Ojala, M. Pietikäinen, and D. Harwood, "A Comparative Study of Texture Measures with Classification Based on Featured Distributions," *Pattern Recognition,* vol. 29, no. 1, pp. 51-59, 1996.

[20] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning,* vol. 20, no. 3, pp. 273-297, 1995.

[21] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," http://www.csie.ntu.edu.tw/~cjlin/libsvm/, 2001.

[22] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 7, pp. 971-987, July 2002.

[23] H.-C. Lian and B.-L. Lu, "Multi-View Gender Classification Using Local Binary Patterns and Support Vector Machines," *Proc. Third Int'l Symp. Neural Networks,* vol. 2, pp. 202-209, 2006.

[24] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 511-518, 2001.

[25] O. Jesorsky, K.J. Kirchberg, and R. Frischholz, "Robust Face Detection Using the Hausdorff Distance," *Proc. Third Int'l Conf. Audio and Video-Based Biometric Person Authentication,* pp. 90-95, 2001.

[26] Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariéthoz, "Measuring the Performance of Face Localization Systems," *Image and Vision Computing,* vol. 24, no. 8, pp. 882-893, Aug. 2006.