Using Component Features for Face Recognition

Yuri Ivanov Bernd Heisele Thomas Serre
Honda Research Institute, US
145 Tremont St
Boston, MA 02111

Abstract

In this paper we explore different strategies for classifier combination within the framework of component-based face recognition. In our current system, the gray values of facial components are concatenated to a single feature vector which is then fed into the face recognition classifier. As an alternative, we suggest to train recognition classifiers on each of the components separately and then combine their outputs using the following three strategies: voting, sum of outputs, and product of outputs. We also propose a novel Bayesian method which weights the classifier outputs prior to their combination. In experiments on two face databases, we evaluate the different strategies and compare them to our existing recognition system.

1 Introduction

The problem of face recognition has been one of the most prominent areas of machine vision for about a decade. Current systems have advanced to be fairly accurate in recognition under constrained scenarios, but extrinsic imaging parameters such as pose, illumination, and facial expression still cause much difficulty in correct recognition.

Recently, component-based approaches have shown promising results in various object detection and recognition tasks such as face detection [11, 15], person detection [9], and face recognition [2, 14]. The component-based face detector described in [4] localizes a set of facial components using a two level hierarchy of classifiers. On top of this detector, we built a component-based face identification system [3] in which the gray values of the extracted components were combined and then classified by a set of Support Vector Machines, one for each person in the database. In experiments, we have shown that the component-based system consistently outperforms holistic face recognition systems in which classification was based on the whole face pattern.

In this paper we investigate alternative techniques for combining components for face recognition. Instead of concatenating the gray values of the extracted components to a single feature vector, we train individual recognition classifiers on each extracted component and merge their outputs using popular combination strategies (cf. [8, 12]) as well as a novel Bayesian approximation method which performs a per-class classifier weighting. Compared to concatenating components, training separate component classifiers is more consistent with the approach chosen in the face detection module where component detectors feed their outputs to a combination classifier.

2 Approach

Our approach to classifier combination is based on viewing the output of each of the multi-class classifiers as a random variable, $\tilde{\omega}$, which takes values from 1 to K, the number of classes. In [6] we propose a Bayesian framework for classifier combination, where outputs of C individual component classifiers, λ_i are weighted by a confidence measure imposed on the classifier performance, $P(\lambda_i|x)$. The probability of the true class label, ω, for a given observation can be approximated by using the empirical error distribution derived from the classifier confusion matrix:

$$\begin{aligned} P(\omega|x) &= \sum_{i=1}^{C}\sum_{k=1}^{K} P(\omega|\tilde{\omega}_k, x, \lambda_i) P(\tilde{\omega}_k|x, \lambda_i) P(\lambda_i|x) \\ &\approx \sum_{i=1}^{C}\left[\sum_{k=1}^{K} P(\omega|\tilde{\omega}_k, \lambda_i) P(\tilde{\omega}_k|x, \lambda_i)\right] P(\lambda_i|x) \end{aligned} \tag{1}$$

The essence of equation 1 is that the prediction of each classifier is weighted in accordance to the error distribution over the predicted class. In the last line of this equation the conditional error distribution, $P(\omega|\tilde{\omega}, x, \lambda_i)$, which is difficult to obtain, is approximated by its projection, $P(\omega|\tilde{\omega}, \lambda_i)$. The latter is simply an empirical distribution that we obtain from the confusion matrix of the classifier on a validation subset of the training data.

This model establishes a general framework for classifier combination, from which a variety of different combination strategies can be derived. In particular, Tax *et. al.*,

[12], present a framework in which sum and product rules are formally justified. Our framework is fully compliant with their work by allowing to justify critic-based (induced by $P(\lambda|x)$) and error-corrected (induced by $P(\omega|\tilde{\omega}, x, \lambda)$) variants of popular combination schemes.

3 Experimental Evaluation

We evaluate four different strategies for face classification on two data sets. First strategy is to use one of the traditional face classification techniques, such as a Support Vector Machine, [13], to classify the gray scale image of the face. The second strategy is to extract components of the face and perform the classification only on the pixels that are part of these components, [3] . The third strategy is to classify each of the face components independently and then combine the results of the classification with one of the standard schemes - *voting*, *product* or *sum*. Lastly, we apply the error-based classifier score correction to the outputs of the classifiers and combine them with the same combination strategies.

3.1 Classifier Score Combination

The techinque, proposed in section 2 presumes that class probabilities are available from each classifier. This presents a minor problem when using discriminative models, such as SVMs. In its direct formulation SVM does not output probabilities, but rather, values of the discriminant function. We convert these scores to probabilities by applying to them the softmax function:

$$P(\tilde{\omega}|x) = \frac{\exp(s_{\tilde{\omega}})}{\sum_{\tilde{\omega}} \exp(s_{\tilde{\omega}})} \tag{2}$$

Using this transformation does not change the classification decision for a minimum error rate classifier, but allows us to treat the classifier within the probabilistic framework of section 2.

3.2 Experiment Setup

In order to evaluate the combination schemes on each data set we train several classifiers:

1. Full face classifier
   A single feature classifier that uses gray-level values of an extracted face image;

2. Component-based stacked face classifier
   A single classifier that has a feature vector formed from all extracted face components placed in the same feature vector;

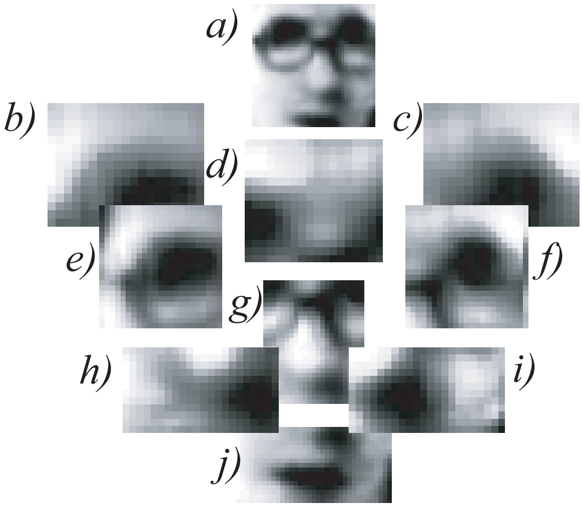

Figure 1: Example of extracted face components. a) bounding box; b,c) eyebrows; d) bridge of the nose; e,f) eyes; g) nose; h,i) nostrils; j) mouth.

3. A set of ten individual component classifiers
   Each extracted face feature is classified independently of others in the set. The resulting outputs are then combined with one of three combination strategies - *voting*, *product* or *sum*;

4. A set of ten individual weighted component classifiers
   The training data used to train the above classifiers is split $90\% - 10\%$. Each classifier is trained on the 90% subset of the data. Then empirical error distribution is computed on the remaining 10%. After the test data is classified the scores are weighted by the resulting distribution (eqn. 1). The resulting outputs are then combined with one of three combination strategies - *voting*, *product* or *sum*.

All classifiers in our experiments are Support Vector Machines with a polynomial kernel of degree 2. To train and test our classifiers we use the SVMFu package [10].

3.3 Results

We evaluate the combination schemes on two data sets. One is a set of faces of six people collected directly from a surveillance system, [6]. The other data set is a combination of synthetic and real faces with synthetic faces used for training. The set includes ten subjects.

From each image in both sets we automatically extract face regions. The detected face images are subsequently histogram-equalized and re-scaled to be 70×70 pixels in size. From these images we extract ten face components, such as eyes, eyebrows, bridge of the nose, mouth, nose, left

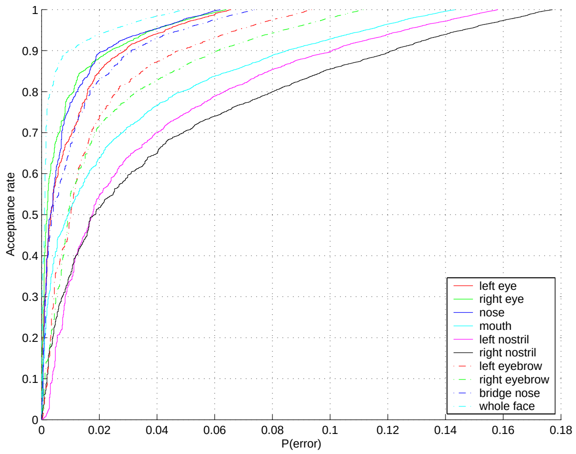

Figure 2: ROC curves for component classifiers for the surveillance data.

and right nostrils, as well as the image area inside the box bounding all of these components. For complete details of the component extraction the reader is referred to [3] . The extracted face components are shown in figure 1.

3.3.1 Surveillance data

The first data set for our experiments is collected from an automated surveillance system. We selected a set of six people recorded within the surveillance area over the course of several days under different natural lighting conditions. The users were not deliberately posed to get a good view of the face, so, the collected data contains only a several thousand faces.

Additionally, due to pecularities of people's behavior around the video camera, some classes got disproportionally few examples, as compared to others. This resulted in a slightly misbalanced data set, however we intentionally left it intact, as it reflects the real situation in which the classifier is to be applied.

All training and testing data are extracted from the images of people automatically with no manual intervention [4] [1]. Some examples of the training and test set are shown in figure 3.

The results of running the experiments on the surveillance data are shown in figures 2 and 4. Figure 2 shows the ROC curves for each component classifier. The performance of the components varies in a wide range, at the top are the whole face, the eyes and the nose classifiers. The nostrils are the least reliable components. The second figure shows a plot of the log-probability of error as a function of acceptance rate. This plot is related to an ROC curve, calculated for a multi-class problem. Given the set of classifier scores (posterior probabilities) we vary a threshold within the full range of posteriors rejecting samples for which the scores fall below the threshold. We compute the error rate for the remaining points and plot it against their fraction in the data set.

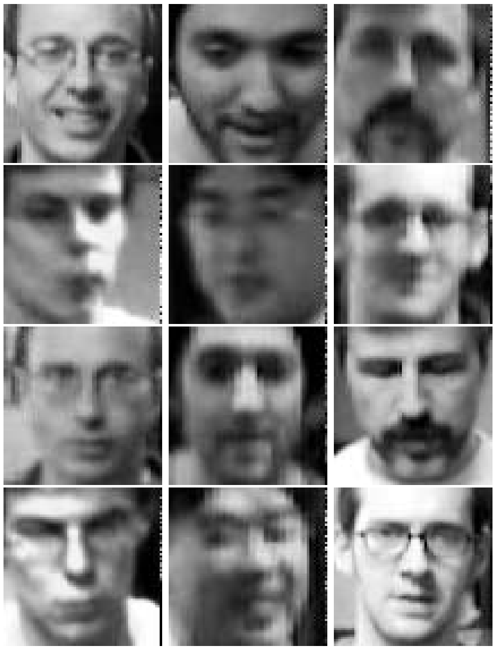

Figure 3: Examples from the surveillance training set (top two rows) and test set (bottom rows). Both sets include mostly frontal views of faces. The resolution of the images was relatively low.

The behavior of the classifiers is illustrated in figure 4. For all acceptance rates, the full face classifier shows the worst performance. The classifier based on stacked features outperforms the full face classifier by a large margin. In turn, all combination schemes consistently provide even better performance. There is only slight difference in performance of weighted and unweighted combination schemes. The latter is explained by the fact that the individual classifiers are very strong on this data set and their confusion matrices are very close to identity, and, hence, have little effect on the combined scores. The difference becomes more pronounced when the strength of the classifiers in the sets varies by a larger factor, as our further experiments show, [6, 7].

3.3.2 Synthetic training data

The second data set is a a combination of synthetic and real human faces. Synthetic faces are used to train classifiers, while real faces are used to test them. In order to generate

[1] This implies that false positives of the face detection phase are left in both training and test data sets, slightly degrading the classifier performance.

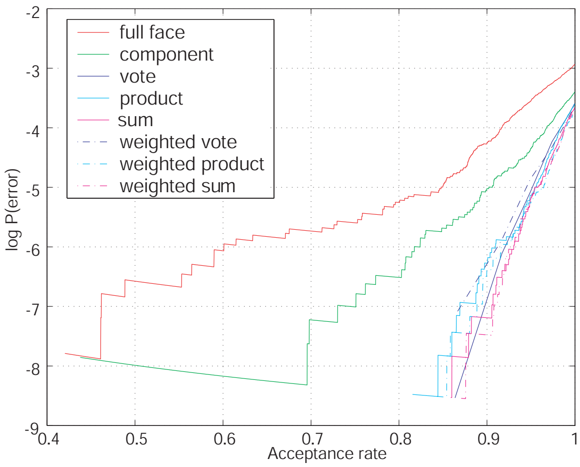

Figure 4: Log-error-acceptance curves for different classifiers for the surveillance data.

the training set we fitted a morphable 3D head model [1, 5] to a frontal and a profile image of each person and then generated a large number of synthetic images by carefully varying the illumination and viewing angles. Overall we generated about 1900 synthetic training images for each of the 10 people in the database.

The test set was created by taking images of the 10 people in the database with a digital video camera. The subjects were asked to rotate their faces in depth and the lighting conditions were changed by moving a light source around the subject. The final test set consisted of about 200 images of each person recorded under different viewpoints and lighting conditions. Some example images of the training and test set are shown in figure 5.

The results of running the experiments are shown in figure 6 for the individual component classifier and in figure 7 for the combinations, respectively. In both cases, the recognition rates are lower than in the previous experiments on the surveillance data. This is not surprising since we only used synthetic images for training. Adding 10% of the test set to the training data and evaluating the retrained systems on the remaining 90% of the original test set lead to a large improvement in the recognition rates. In addition, we ran experiments where we recomputed the weights of the classifiers on 10% of the test set but did not retrain the component classifiers. The results are shown in figure 8. As expected, the best performance is achieved by retraining the classifiers on the enlarged training set including real face images. Interestingly, recomputing the weights while keeping the component classifiers unchanged also leads to a clear improvement compared to the original systems. This indicates that our weighting algorithm can be used to perform on-line learning in applications in which retraining of all classifiers is too time consuming or not possible at all.

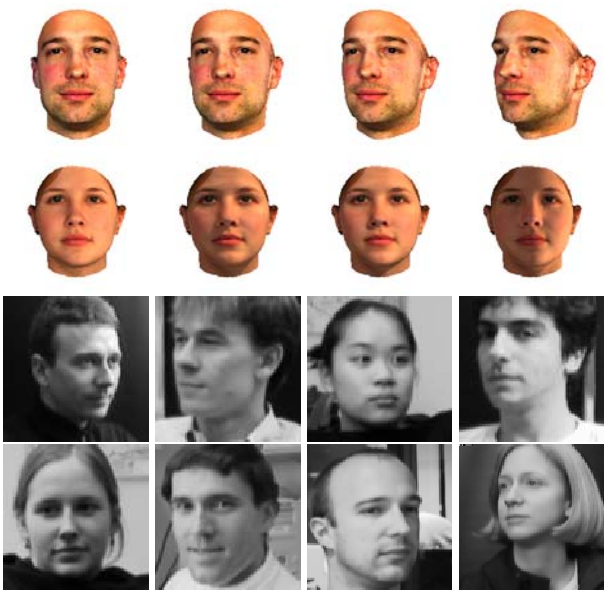

Figure 5: Examples from the synthetic training set (top rows) and test set (bottom rows). Note the variety of poses and illumination conditions.

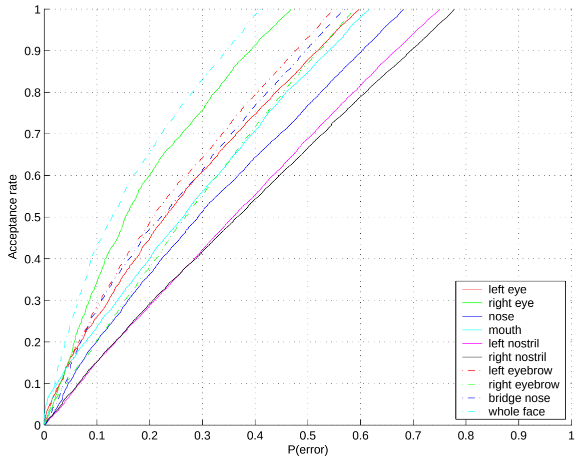

Figure 6: ROC curves for component classifiers trained on synthetic images.

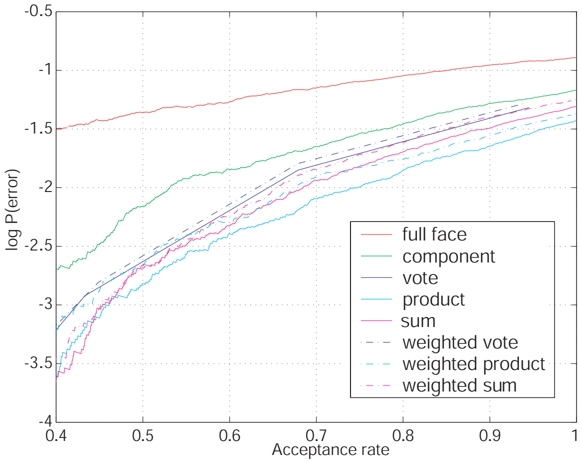

Figure 7: Log-error-acceptance curves for different classifiers trained on synthetic images.

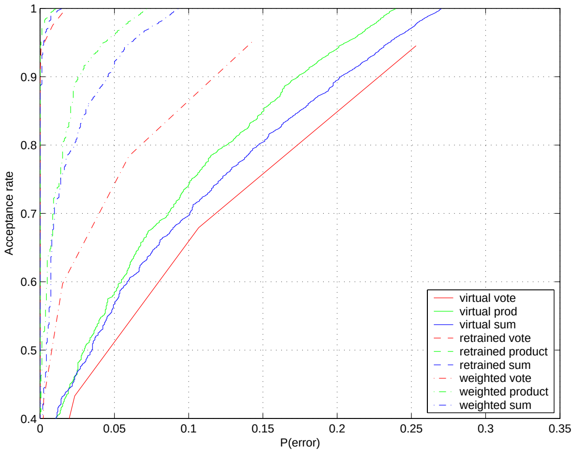

Figure 8: ROC curves for Classifiers trained only on synthetic data, using part of the test data without retraining and with full retraining.

4 Conclusion

In our previous face recognition system facial components were extracted from the input image and combined into a single feature vector which was then fed into the recognition classifier. As an alternative, we proposed to train recognition classifiers on each of the components separately and then combine their outputs. Three popular strategies for combining the outputs have been evaluated: voting, sum of outputs, and product of outputs. We also proposed a new method based on the distribution of the empirical error for weighting the outputs prior to their combination.

Experiments were carried out on two data sets: the first set included mostly low resolution, frontal face images of six people recorded by a surveillance camera. The second set consisted of synthetic training images of ten people and real test images with large variations in pose and illumination. On both sets we achieved a significant improvement over our previous system. Overall, the product of the outputs marginally outperformed the other two combination strategies. By running our new weighting algorithm on 10% of the test set, we could increase the recognition rate by a large margin without having to retrain the component classifiers.

References

[1] V. Blanz and T. Vetter. A morphable model for synthesis of 3D faces. In *Computer Graphics Proceedings SIGGRAPH*, pages 187–194, Los Angeles, 1999.

[2] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, 1993.

[3] B. Heisele, P. Ho, and T. Poggio. Face recognition with support vector machines: Global versus component-based approach. In *International Conference on Computer Vision*, pages 688–694, Vancouver, Canada, 2001.

[4] B. Heisele, T. Serre, M. Pontil, T. Vetter, and T. Poggio. Categorization by learning and combining object parts. In *Advances in Neural Information Processing Systems*, volume 14, 2002.

[5] J. Huang, B. Heisele, and Blanz V. Component-based face recognition with 3D morphable models. In *4th Conference on Audio- and Video-Based Biometric Person Authetication*, 2003.

[6] Y. Ivanov, T. Serre, and J. Bouvrie. Error-weighted classifier combination for multi-model human identification. In *Submission*, 2004.

[7] A. Kapoor, R. W. Picard, and Y. Ivanov. Probabilistic combination of multiple modalities to detect interest. In *Submission*, 2004.

[8] J. Kittler, Y. P. Li, J. Matas, and M. U. Ramos Sánchez. Combining evidence in multimodal personal identity recog-

nition systems. In *Intl . Conference on Audio- and Video-Based Biometric Authentication*, Crans Montana, Switzerland, 1997.

[9] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 23, pages 349–361, April 2001.

[10] Ryan Rifkin. SVMFu package. http://five-percent-nation.mit.edu/SvmFu/index.html, 2000.

[11] H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 746–751, 2000.

[12] David M. J. Tax, Martijn Van Breukelen, Robert P. W. Duin, and Josef Kittler. Combining multiple classifiers by averaging or by multiplying? *Pattern Recognition*, 33:1475 – 1485, 2000.

[13] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, Berlin, 1995.

[14] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

[15] D. Xi and S.-W. Lee. Face detection and facial component extraction by wavelet decomposition and support vector machines. In *4th International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA)*, pages 199–207, 2003.