International Journal of Computer Vision 74(2), 167–181, 2007

DOI: 10.1007/s11263-006-0006-z

A Component-based Framework for Face Detection and Identification

BERND HEISELE
Honda Research Institute USA in Cambridge and the Center for Biological and Computational Learning at M.I.T. Cambridge

THOMAS SERRE AND TOMASO POGGIO
McGovern Institute for Brain Research, the Center for Biological and Computational Learning, the Department of Brain and Cognitive Sciences and the Computer Science and Artificial Intelligence Laboratory at M.I.T.

Received May 1, 2006; Accepted November 6, 2006

First online version published in December, 2006

Abstract. We present a component-based framework for face detection and identification. The face detection and identification modules share the same hierarchical architecture. They both consist of two layers of classifiers, a layer with a set of component classifiers and a layer with a single combination classifier. The component classifiers independently detect/identify facial parts in the image. Their outputs are passed the combination classifier which performs the final detection/identification of the face.

We describe an algorithm which automatically learns two separate sets of facial components for the detection and identification tasks. In experiments we compare the detection and identification systems to standard global approaches. The experimental results clearly show that our component-based approach is superior to global approaches.

Keywords: face detection, face identification, face recognition, object detection, object recognition, support vector machines, components, fragments, parts, hierarchical classification

1. Introduction

Object detection systems in which classification is based on local object features have become increasingly common in the computer vision community over the last couple of years (see e.g. Ullman et al., 2002; Heisele et al., 2001; Mohan et al., 2001; Weber et al., 2000; Dorko and Schmid, 2003; Schneiderman and Kanade, 2000). These systems have the following two processing steps in common: In a first step, the image is scanned for a set of characteristic features of the object. For example, in a car detection system a canonical gray-value template of a wheel might be cross-correlated with the input image to localize the wheels of a car. We will refer to these local object features as the components of an object, other authors use different denotations such as parts, patches or fragments. Accordingly, the feature detectors will be called component detectors or component classifiers. In a second step, the results of the component detector stage are combined to determine whether the input image contains an object of the given class. We will refer to this classifier as the combination classifier.

An alternative approach to object classification is to search for the object as a whole, for example by computing the cross-correlation between a template of the object and the input image. In contrast to the component-based approach, a single classifier takes as input a feature vector containing information about the whole object. We will refer to this category of techniques as the global approach; examples of global face detection systems are described in Sung (1996), Oren et al. (1997), Rowley et al. (1998), Osuna (1998), Heisele et al. (2003). There are systems which fall in between the component-based and the global approach. The face detection system in Viola and Jones (2004), for example, performs classification with an ensemble of simple classifiers, each one operating on locally computed image features, similar to component detectors. However, each of these simple classifiers is only applied to a fixed x-y-position within the object window. In the component-based approach described in

this paper, the locations of the components relative to each other are not fixed; each component detector performs a search over some part of the image to find the best matching component.

In the following we briefly motivate the component-based approach:

(a) A major problem in detection is the variation in the appearance of objects belonging to the same class. For example, a car detector should be able to detect SUVs as well as sports cars, even though they significantly differ in their shapes. Building a detector based on components which are visually similar across all objects of the class might solve this problem. In the case of cars, these indicator components could be the wheels, the headlights or the taillights.
(b) Components usually vary less under pose changes than the image pattern of the whole object. Assuming that sufficiently small components correspond to planar patches on the 3D surface of the object, changes in the viewpoint of an object can be modelled as affine transformations on the component level. Under this assumption, view invariance can be achieved by using affine invariant image features in the component detector stage as proposed in Dorko and Schmid (2003). A possibility to achieve view invariance in the global approach is to train a set of view-tuned, global classifiers as suggested by Poggio and Edelman (1990). However, it is preferable to achieve view invariance at the feature level in order to keep the number of required training examples small. Furthermore, shifting a set of view-tuned classifiers over the image significantly increases the computation at run-time.
(c) Another source of variations in an object's appearance is partial occlusion. In general it is difficult to collect a training set of images which covers the spectrum of possible variations caused by occlusion. In the component-based approach, partial occlusion will only affect the outputs of a few component detectors at a time. Therefore, a solution to the occlusion problem might be a combination classifier which is robust against changes in a small number of its input features, e.g. a voting-based classifier. Another possibility is to add artificial examples of partially occluded objects to the training data of the combination classifier, e.g. by decreasing the component detector outputs computed on occlusion-free examples. Experiments on detecting partially occluded pedestrians with a component-based system similar to the one describe in our paper have been reported in Mohan et al. (2001).

One of the main problems that has to be addressed in the component-based approach is how to choose a suitable set of components. A manually selected set of five components containing the head, the upper body, both arms, and the lower body has been used in Mohan et al. (2001) for person detection. Although there are intuitively obvious choices of components for many types of objects, such as the eyes, the nose and the mouth for faces, a more systematic approach is to automatically select the components based on their discriminative power. In Ullman et al. (2002) components of various sizes were cropped at random locations in the training images of an object. The mutual information between the occurrence of a component in a training image and the class label of the image was used as a measure to rank and select components. An alternative to ranking the randomly extracted components is to cluster them and to use the cluster centers as canonical component templates, as suggested in Morgenstern and Heisele (2003). Another strategy to automatically determine an initial set of components is to apply a generic interest operator to the training images and to select components located in the vicinity of the detected points of interest (Fergus et al., 2003; Dorko and Schmid, 2003; Lowe, 2004). In Dorko and Schmid (2003), this initial set was subsequently reduced by selecting components based on mutual information and likelihood ratio. Using interest operators has the advantage of being able to quickly and reliably locate component candidates in a given input image. However, forcing the locations of the components to coincide with the points detected by the interest operator considerably restricts the choice of possible components—important components might be lost. Furthermore, interest operators have a tendency to fail for objects with little texture and objects at a low pixel resolution. In this paper, we propose a method for automatically learning components for detection and identification based on a training and cross-validation set of faces. This method is a modification of the algorithm proposed in Heisele et al. (2001), in which the size and shape of facial components was learned by minimizing a bound on the classification error of the component classifiers.

How to include information about the spatial relationship between components is another important question that has to be addressed in the component-based approach. In the following discussion we assume that scale and translation invariance are achieved by sliding a window over the input image at different resolutions—the detection task is then reduced to classifying the pattern within the current window. Intuitively, information about the location of the components is important in cases where the number of components is small and each component carries only little class-specific information. Omitting any spatial information leads to a detection system similar to the biological object recognition models proposed in Riesenhuber and Poggio (1999); Ullman et al. (2002); Serre et al. (2005). In Riesenhuber

and Poggio (1999), the components are located by searching for the maximum outputs of the detectors across the window. The only data propagated to the combination classifier are the outputs of the component detectors while the information about the location of the maxima is not used. In this paper, we will investigate two possibilities to add spatial information to this basic system: (a) implicitly by restricting the location of each component to be within a pre-defined search region inside the window, and (b) explicitly by adding the position of the detected components as additional inputs to the combination classifier. An iterative search for facial components based on the detector outputs and the empirical distribution of the relative position between pairs of components has been published in Bileschi and Heisele (2002). This technique achieved a higher accuracy in localizing the eyes in a face than the straightforward search for the maximum detector responses. However, its computational complexity is a major drawback regarding real-world applicability.

The tasks of face detection and identification share the key problems of pose and illumination invariance. Most of the arguments mentioned before in support of component-based face detection can therefore be applied to component-based face identification. However, face identification poses two additional problems: the large number of classes and the small number of training examples per class (Phillips, 1998). Our learning-based approach requires a large set of training examples per person. To solve this dilemma, we apply 3D morphable models (Blanz and Vetter, 1999) during training to generate a sufficient number of synthetic face images from an initially small training set of real images. We do not address the problem of identifying a large number of people.[1] Instead, we focus on robustness against changes in pose and illumination given a small group of people. Systems like ours might be used in home and office environments where the number of people to be recognized is usually small.

In the following, we briefly review face identification techniques which are closely related to our approach. A comprehensive survey on state-of-the-art techniques in face identification can be found in Zhao et al. (2003). In Brunelli and Poggio (1993), faces were identified by independently matching templates of three facial regions: both eyes, the nose and the mouth. The configuration of the components during classification was unconstrained since the system did not include a geometrical model of the face. A similar approach with an additional alignment stage was proposed in Beymer (1993). In an effort to enhance the robustness against pose changes the originally global eigenface method has been further developed into a component-based system in Pentland et al. (1994) where PCA is applied to local facial components. The elastic grid matching algorithm described in Wiskott et al. (1997) uses Gabor wavelets to extract features at grid points and graph matching for the proper positioning of the grid. The identification is based on wavelet coefficients that were computed on the nodes of a 2D elastic graph. In Nefian and Hayes (1999), a window was shifted over the face image and the DCT coefficients computed within the window were fed to a 2D hidden Markov model. A probabilistic approach using part-based matching has been proposed in Martinez (2002) for expression invariant and occlusion tolerant recognition of frontal faces.

Our face identification system uses the outputs of the component-based face detector during training and at runtime. In the training stage, the computed locations of the face detection components are used to iteratively learn a set of components suitable for identification, called identification components. At runtime, the face detector supplies the identification module with the center locations of the identification components which are then extracted and classified by a hierarchy of identification classifiers. This is different from our previous face identification system (Heisele et al., 2003) in which the gray values of the detection components were combined into one feature vector and then classified by a single classifier.

The outline of the paper is as follows: Section 2 explains the architecture of our component-based face detection and identification system. In Section 3 we describe an algorithm for learning facial components for detection and identification. In Section 4 we explore different techniques of integrating spatial information into the detection process and present experimental results on the face detection system. Results of the full system on a face identification test set are given in Section 5. This section also includes comparisons to standard global approaches based on SVM, PCA and LDA.

2. Architecture of the System

An overview of our system is shown in Fig. 1. We first computed a resolution pyramid from the input image. The pyramid was scanned for faces by sliding a fixed sized object window pixel-by-pixel across each image. On the first level of the detection module, component classifiers independently located components of the face inside the current object window. We experimented with two different strategies for localizing the components: (a) searching for the maximum real-valued output of the corresponding component classifier over the whole object window, and (b) searching for the maximum output in a pre-defined rectangular search region within the object window.

The component detectors were linear SVMs, each of which was trained on a set of extracted facial components and on a set of non-face patterns.

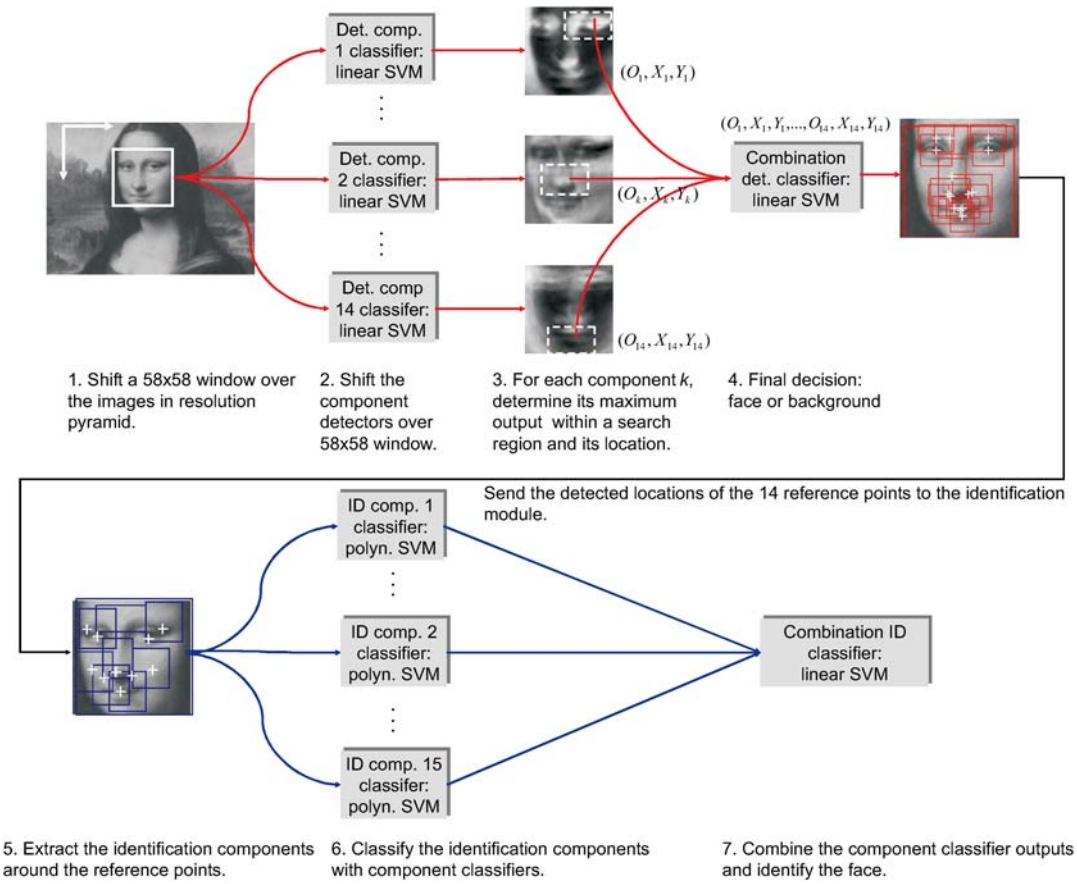

Figure 1. System overview of the component-based face detection and identification system. An object window of fixed size is slid over the input image (1). At the first level, component classifiers specialized in detecting facial parts are shifted over the current object window (2) and their real-valued output values are computed at each position inside the object window (3). At the second level, the maximum output O of each component and optionally the position X, Y of each maximum are input to the detection combination classifier. When a face is detected the center coordinates of the detected components are passed to the identification module and the set of identification components is extracted around these points (5). The identification components are classified separately by second-degree polynomial SVMs (6) and the outputs of the classifiers are combined to identify the person's face (7).

If a face was detected inside an object window, the scale of the face and the location of the detected components were passed to the face identification module. Based on this information, the face identification module extracted a previously learned set of identification components which was different from the set of detection components. Each identification component is classified by a separate component classifier and their results are passed to a combination classifier which performs the final identification of the face. The component classifiers were second-degree polynomial SVMs which were trained on a set of synthetic face images. We implemented four types of combination classifiers: classification based on the majority vote amongst the component classifiers, a classifier based on the sum over the real-valued component classifier outputs, a classifier based on the product of the real-valued component classifier outputs, and a linear SVM trained on the component classifier outputs.

3. Learning Components for Face Detection and Identification

Extracting and labelling training data is usually a tedious and time-consuming work. In order to train the component classifiers for both identification and detection, we have to extract the components from each face image in the training database. Manual extraction would only be feasible for a very small number of components and face images. To automate the extraction process, we used textured 3D head models (Vetter, 1998) with known 3D correspondences. By rendering the 3D head models we could generate faces in arbitrary poses and with arbitrary illumination.

3.1. Training Data for Detection

From 100 textured 3D head models of Caucasian subjects we rendered[2] tens of thousands of face images of

Figure 2. Examples of synthetic face images which were generated for training the component detectors. The size of each image was 100×100 pixels corresponding to resolution of the face of about 58×58 pixels.

Figure 3. Examples of the image triplets used for generating the 3D models.

size 58 × 58 from which we randomly selected 6,843 images for training and cross-validation (CV). Some examples of synthetic face images from the set used to train the face detection system are shown in Fig. 2. The negative training images initially consisted of 10,209 58 × 58 non-face patches randomly extracted from a set of non-face images. We then applied bootstrapping to enlarge the training data by non-face patterns that look similar to faces. To do so, we trained a single linear SVM classifier and applied it to the previously used set of non-face images. The false positives (FPs) were added to the non-face training data to build the final training set of size 13,654. We set one third of the positive and negative images aside to build a CV set of 2,281 face and 4,452 non-face patterns. This left 4,562 face and 9,102 non-face patterns for training.

3.2. *Training Data for Identification*

We first fitted 3D face models to three images of each person in the face identification database (Blanz and Vetter, 1999). Examples of the image triplets are shown in Fig. 3. Each triplet consisted of a frontal, a half-profile, and a profile high resolution face image. We then generated synthetic faces at a resolution of 58 × 58 for the ten subjects by rendering the 3D face models under varying pose and illumination. The original frontal face images of all ten subjects and the corresponding synthetic images are shown in Fig. 4. The images were divided into a training set of 3,080 images and a CV set of 3,960 images.[3]

3.3. *Learning of the Components*

The learning algorithm iteratively grew components around a manually preselected set of points in the face image, called reference points (see Fig. 5). The same set of 14 reference points was used to learn the detection and the identification components. Choosing the same reference points was crucial since it allowed us to use the face detector to localize both types of components at runtime. The learning algorithm described in the following paragraphs was applied to each component separately.

The growing algorithm started with a small rectangular component located around a reference point in the face image. For learning the detection components, the position of each reference point was accurately determined based on the 3D correspondences given by the morphable model. For the identification components, we ran the face detector on the training and CV images of the identification to locate the reference points in each image. The identification components were then extracted around the reference points from all face images to build a training set of positive examples. For detection, the negative set consisted of random background patterns which had the same size and rectangular shape as the facial component. As component classifier we chose a linear SVM. In the case of identification, each component classifier was a set of second-degree polynomial SVMs which were trained according to the one-vs.-all strategy, i.e. the components of one person were trained against the components of the remaining nine people. For both detection and identification the components were histogram-equalized to remove variations caused by lighting changes.

After training a component classifier on the histogram-equalized gray values of the extracted components we determined its performance on the CV set.[4] We then enlarged the component by expanding the rectangle by one pixel into one of the four directions: up, down, left, and right. As before, we generated the training data, trained a component classifier and determined its error rate on the CV set. We did this for expansions into all four directions and finally kept the expansion which led to the smallest error. This process was continued until the error on the CV set reached zero or until a maximum number of iterations had been computed. The iterative growing process of the

Figure 4. Original images and synthetic images generated from 3D models for all ten subjects in the identification training database.

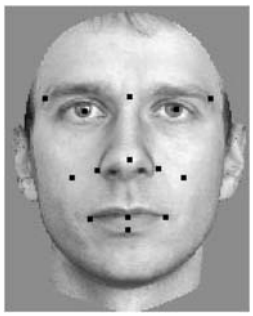

Figure 5. The 14 reference points used for component learning.

right eye and nose components for the detection case is shown in Fig. 6. The final set of components is depicted in Fig. 7 and their dimensions are given in Table 1.

4. Evaluation of the Face Detector

4.1. Spatial Information for Face Detection

As mentioned in the introduction, we experimented with two different strategies for localizing the components:

(a) Searching for the maximum real-valued output of each component classifier over the whole object window.
(b) Searching for the maximum output within a rectangular search region around the expected location of the component. The search region was computed as the smallest rectangle that contained the position of the center point of the given component across all training images.

The maximum output of a component classifier computed across the whole pattern of a face is likely to be close to the expected location of the component, i.e. to fall within the search region of the component computed on the training data. For a non-face pattern, on the other hand, the component classifier's maximum can be assumed to be uniformly distributed across the whole pattern. Therefore, the maxima computed within search regions can be expected to have a higher discriminative power than the maxima computed across the whole pattern. This argument loses in strength as the variations in the pose increase which in turn leads to an increase in the size of the search regions. Another weakness of modelling the geometry of an object using the maximum operation within search regions is that it does not account for the correlation between the positions of the components in the detection process since each component is detected separately within its corresponding search region. A way of exploiting the pairwise correlation between the positions of components during the detection stage has been suggested in Bileschi and Heisele (2002).

Another way of including spatial information into the classification process is to propagate information about the image location of the detected component to the combination classifier. To analyze this possibility, we

Figure 6. Illustration of the iterative growing process for the component located at the center of nose (top) and the right eye (bottom) in the detection case. The components are arranged from top left to bottom right, starting with the initial component in the top left corner. The minimum CV error is given above each component and the arrow indicates the corresponding direction selected.

Table 1. The dimensions of the 14 learned components for face detection and identification. Left and right are relative to the face.

	Detection				Identification			
Component	Right	Left	Up	Down	Right	Left	Up	Down
R. Eye	9	17	9	9	7	7	18	4
L. Eye	17	9	9	9	19	2	14	3
R. Eyebrow	2	31	5	22	2	12	9	7
L. Eyebrow	31	2	5	22	17	2	12	7
Nose Bridge	7	13	8	8	8	9	17	4
Nose	10	8	19	17	8	3	5	22
R. Nostril	11	13	10	10	9	11	11	7
L. Nostril	13	11	10	10	14	4	12	8
R. Corner Mouth	9	24	19	5	3	8	9	13
L. Corner Mouth	24	9	19	5	16	2	12	8
Upper Lip	10	9	8	5	10	4	10	13
Lower Lip	15	9	22	1	9	3	10	10
R. Cheek	6	31	7	16	3	15	6	14
L. Cheek	31	6	7	16	21	2	5	10

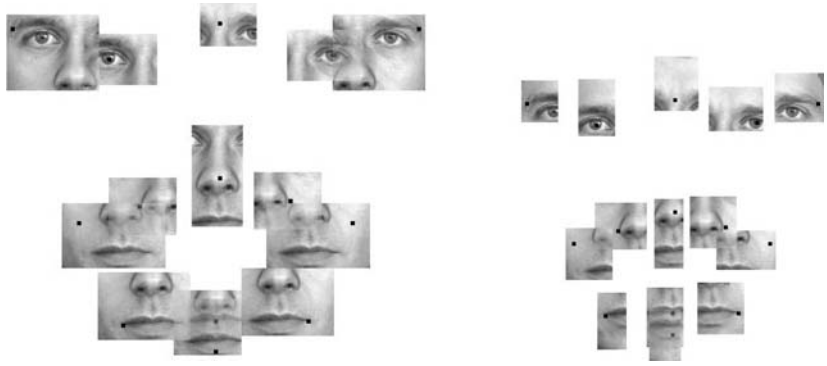

Figure 7. The final sets of the 14 learned components for detection (left) and identification (right).

conducted experiments with three types of feature vectors for the combination classifier:

(1) The maximum real-valued outputs o_i of the component classifiers. This lead to a feature vector $(o_1, \ldots, o_i, \ldots, o_{14})$.
(2) The absolute-valued deviations $(|x_i|, |y_i|)$ from the expected position of the component in the image[5] which lead to a $2N$-dimensional vector $(|x_1|, |y_1|, \ldots |x_i|, |y_i|, \ldots, |x_{14}|, |y_{14}|)$.
(3) The concatenation of the previous two feature vectors: $(o_1, |x_1|, |y_1|, \ldots o_i, |x_i|, |y_i|, \ldots, o_{14}, |x_{14}|, |y_{14}|)$.

4.2. Results

The test set consisted of 5,000 non-face patterns which were selected by a 19×19 low-resolution LDA classifier as the most similar to faces out of 112 background images. The positive test set consisted of a subset of the CMU-PIE database (Sim et al., 2003) which we randomly sampled across the individuals, illumination and expressions. We restricted the rotation of the faces to be in the range between about -30° to 30° which matched the pose range spanned by the training set. The faces were extracted based on the coordinates of facial feature points given in the CMU-PIE database.

For each test image, we computed the real-valued outputs of the combination classifier across different scales and positions.[6] Only the maximum output of the combination classifier in a given test image was kept for computing the ROC curve. Figures 8–10 show examples of

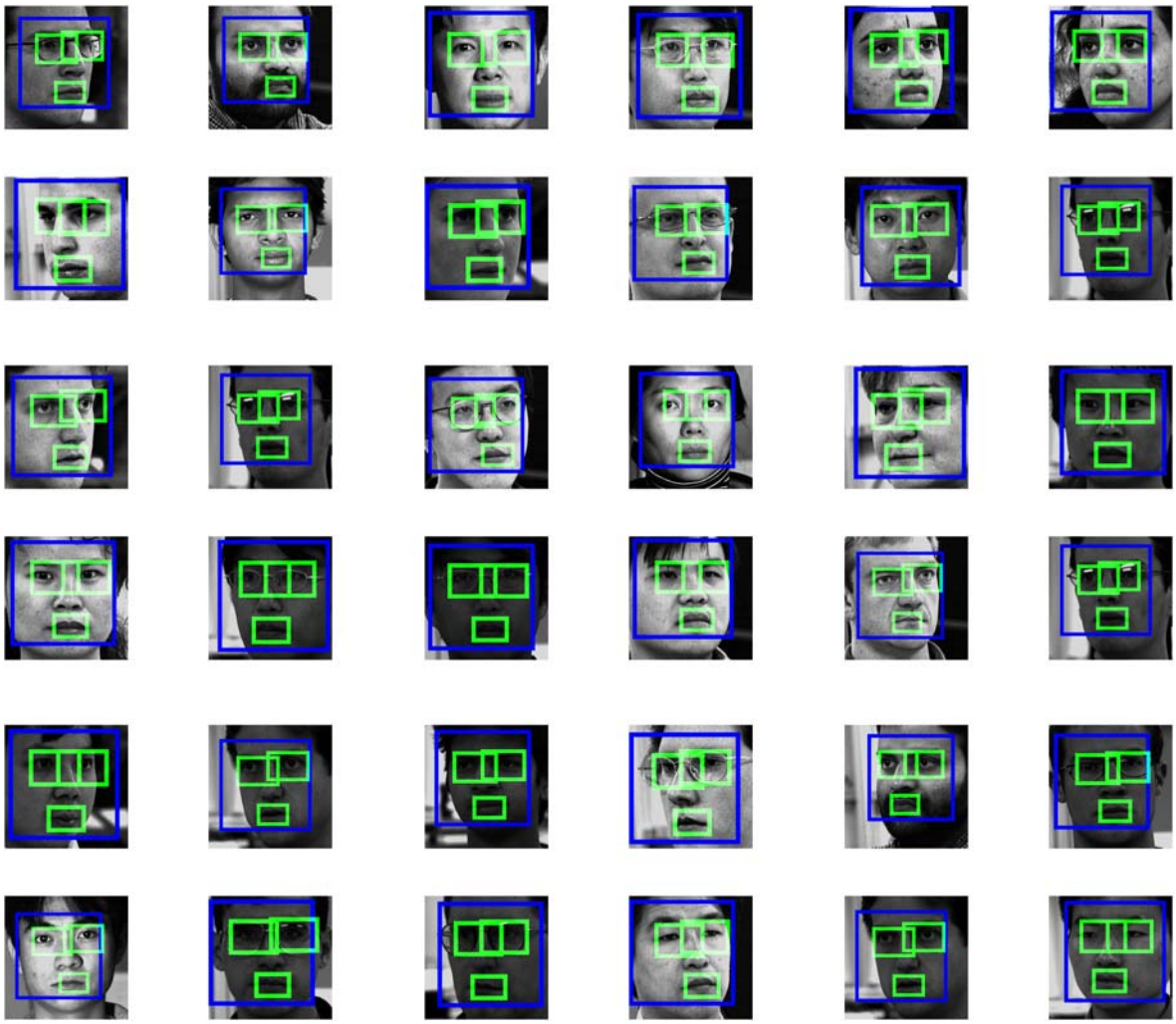

Figure 8. Examples of faces which were correctly detected by the component-based face detector.

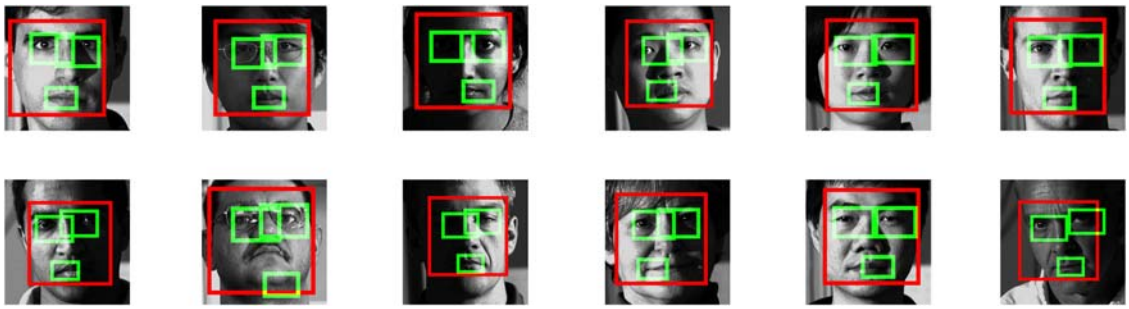

Figure 9. Examples of faces which were missed by the face detector. Errors are mainly due to limitations of the synthetic training set (lacking variations in lighting, absence of facial hair, eye glasses and facial expression).

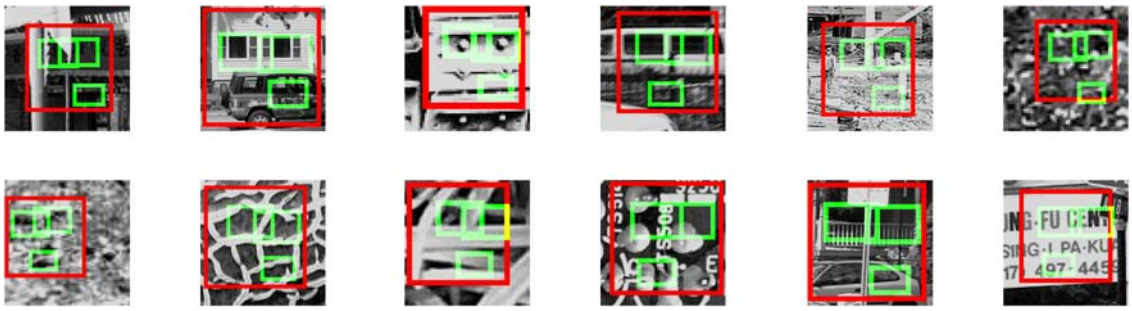

Figure 10. Examples of false detections generated by the component-based face detector.

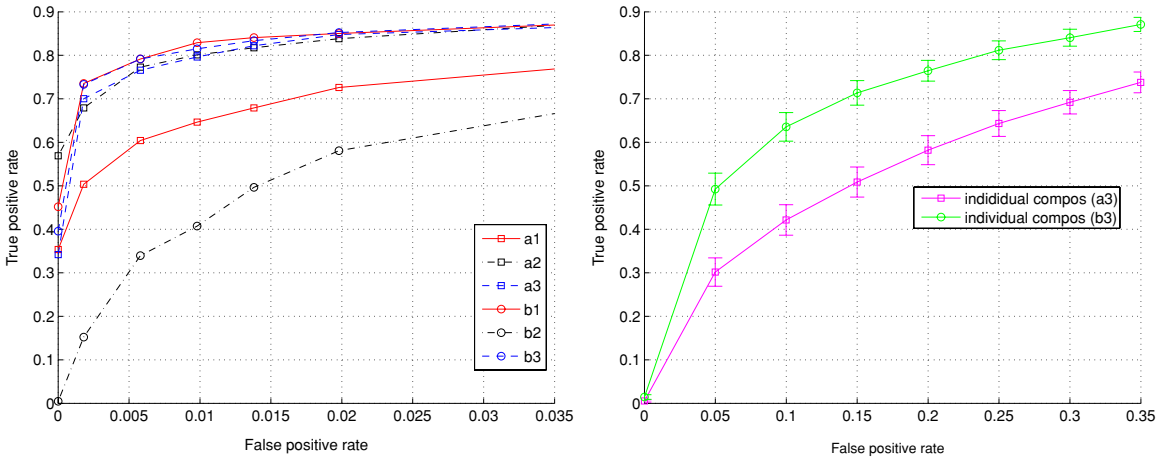

Figure 11. Left: Comparison between different classification strategies: "a" no search regions, "b" search regions, "1" maximum outputs of the component classifiers as inputs to the combination classifier, "2" positions of the detected components as inputs to the combination classifier, "3" positions and maximum outputs as inputs to the combination classifier. Right: Comparison between the performance of the component classifiers for the systems without search regions "a" and with search regions "b". Each of the two ROC curves is the average across the 14 ROC curves of the component classifiers.

faces which were correctly detected, missed, and falsely detected. The red (correct) and blue (incorrect) boxes indicate the most likely locations of a face according to the maximum response of the combination classifier. The smaller green boxes indicate the location of the two eyes and the mouth. Note that the variability in the training images was much smaller than in the test images. The synthetic training database consisted of Caucasian faces only, it did not include people with a beard or a moustache, neither did it include people wearing glasses.

The left diagram in Fig. 11 shows the ROC curves for the different classification strategies. The two systems with search regions and a feature vector containing the maximum outputs of the component classifiers (b1 and b3) perform about the same and are better than the rest. The system with search regions and position features only (b2) performs poorly. It is not surprising that once the search for a component is confined to a relatively small region, the position of the maximum is not a good feature to distinguish between faces and non-faces. The systems in which the components are searched across the whole pattern (a1, a2, a3) perform worse than b1 and b3 but better than b2. The importance of search regions is also evident in the comparison between the individual component classifiers for a3 and b3. The diagram on the right in Fig. 11 shows the ROC curves averaged across the 14 components. The difference in the recognition rate between search regions and no search regions is about 10% across large parts of the ROC curve. Comparing the two diagrams in Fig. 11 we get an impression of the improvement achieved by combining the component classifiers. The FP rate of the classifier combination at 70% recognition rate is about 1% of the FP rate of the individual classifiers, at 80% recognition rate it is about about 2%.

The ROC curve in Fig. 12 compares the best of the component-based systems (b1) with two global systems and the OpenCV (OpenCV Online Reference Manual, 2006; Lienhart et al., 2003) implementation of the Viola and Jones (2004) face detection system. The global classifiers were single SVMs with linear and second-degree polynomial kernels trained on the histogram-equalized gray values of the whole 58×58 face patterns.

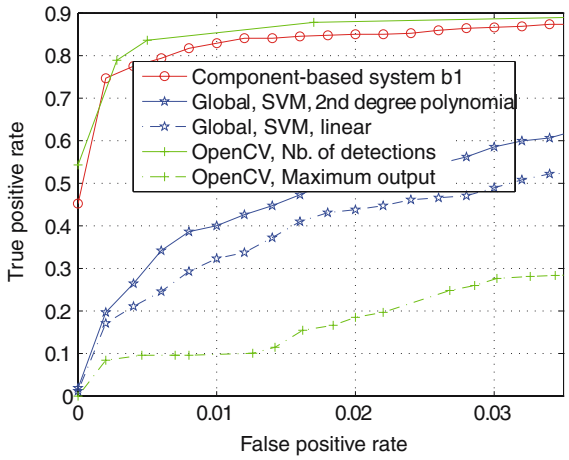

Figure 12. Comparison between the best component-based system ("b1") with two global SVM classifiers (linear and a second-degree polynomial) and the OpenCV frontal face detector based on Viola and Jones (2004). We used two different criteria to compute the ROC curve for the OpenCV system: the maximum output of the final classifier ("Maximum output") and the number of overlapping detections ("Nb. of detections") per test pattern.

The component system clearly outperforms both global systems.

We compared our system to OpenCV's default frontal face detector[7] and we retrained the OpenCV system on our training data. The performance of the retrained system was worse than that of OpenCV's default detector. The recognition rate for 15 stages was 64% with a FP rate of 10%, i.e. 36% of the test faces did not reach the 15th stage. The poor performance of the retrained classifier might be explained by the uniformity of our training data (perfectly aligned, synthetic faces) which could cause overfitting. In Lienhart et al. (2003), small rotations and distortions were applied to the face images to enlarge the training set. Another problem might be the relatively small number of negative training samples —hierarchical classifiers of this type are usually trained with millions of negative samples, see e.g. the experiments in Lienhart et al. (2003) and Viola and Jones (2004). The component-based system performed similar to OpenCV's default frontal face detector, which was trained on real face images and a much larger number of negative training samples. Note that the OpenCV software computed the ROC curve based on the number of detections within a neighborhood, while we computed the ROC curve for our system based on the maximum output of the classifier across a test pattern. The two techniques yield vastly different ROC curves, as can be seen in Fig. 12.[8] If we base the performance comparison on our method of computing the ROC curve, the component-based detector would have a clear advantage over the OpenCV detector.

A short note on the computational costs of our system compared to the OpenCV system: Training of the component system, once the shapes of the components had been fixed, was several orders of magnitude faster than training the OpenCV detector. During classification, our system had to run 14 component detectors across each scaled instance of the original input image. Since we used linear SVMs, this was equivalent to correlating the image with 14 component templates. In addition we had to compute the maxima of the component detector outputs within the search regions. Since the search regions were smaller than the components and since the maximum search can be implemented as a one-dimensional search across the rows and columns of the image, the computational costs of the maximum computation were negligible compared to the costs of the correlation. According to Viola and Jones (2004), a template based system with a single global template of size 24×24 required about 20 times more computations than the hierarchical system in Viola and Jones (2004).

5. Evaluation of the Face Identifier

5.1. *Procedure*

As previously described, the face identification was based on the 14 identification components which were iteratively grown around the same 14 reference points as the detection components. The identification module could therefore use the positions of the detection components

Figure 13. Examples of correct classifications from the identification test set. Shown is a pair of images for each subject in the database. The subjects 1 to 10 are arranged from top left to bottom right. Note the variety in pose and illumination.

Figure 14. Examples of misclassified faces. Upon visual inspection about 50% of the failures could be attributed to inaccurate component detection due to pose, expression and illumination. In the remaining cases there was no obvious reason for the misclassification.

computed by the face detector to localize and extract the identification components. We added a global component to the set of 14 learned components which covered a large part of the face. The location of this component was computed by taking the circumscribing square around the 14 reference points. After extraction, the squared image patch was scaled to a fixed size of 40×40 pixels. We performed histogram equalization on each of the 15 components separately and then input their gray values to the corresponding second-degree polynomial SVM classifier.

For training the 15 component classifiers we used the full set of the 7,040 synthetic face images described previously, i.e. we did not split the set into training and CV sets as was done for learning the identification components. A test set was created by recording images of the ten people in the database with a digital video camera. The subjects were asked to rotate the face in depth and the lighting conditions were changed by moving a light source around the subject. The final test set consisted of 200 images of each person. The training and test images were recorded on different days and with different cameras.[9]

5.2. *Results*

The component-based face identification system was compared to three types of global face identification systems, a PCA-based system, an LDA classifier, and a second-degree polynomial SVM. The input to all global systems was the 58×58 histogram equalized face pattern as extracted by the global face detector (second-degree polynomial SVM) described in the previous section. In the PCA experiment we computed the 300-dimensional eigenspace of all extracted faces from the training set of 7,040 images.[10] During testing, we projected a given face into the 300-dimensional PCA subspace and then computed the closest neighbor amongst the training images.

We ran the component-based face detector (version b1) on the face identification test set to find the face and to determine the positions of the 14 reference points in the face. We then cropped the 14 learned identification components plus the global component from the image, histogram-equalized them and classified them by their corresponding component classifiers. We explored four techniques for combining the outputs of the component classifiers (Ivanov et al., 2004) and compared it to our previous approach in which we stacked the component features into a single feature vector (Heisele et al., 2003):

Table 2. The recognition rates.

Classifier	Correct [%]	Errors [%]
Component-based		
Majority vote	86.80	13.20
Maximum product	88.85	11.15
Maximum sum	87.20	12.80
SVM, linear	**89.25**	**10.75**
SVM, linear, detection components	86.40	13.60
SVM, linear, single global component	77.90	22.10
Stacked features	84.65	15.35
Global		
LDA	61.10	38.90
PCA, 300 dimensions	52.70	47.30
SVM, 2nd degree polynomial	63.40	36.60

- *Majority vote:* We computed the majority vote amongst the discrete-valued component classifier outputs. In case of a tie we decided based on the maximum product.
- *Maximum product:* For each of the ten classes we computed the product of the real-valued classifier outputs and then selected the class with the largest product. Prior to taking the product we normalized each classifier's outputs using the softmax function.
- *Maximum sum:* For each of the ten classes we computed the sum over the real-valued classifier outputs

and then selected the class with the largest sum. Prior to computing the sum we normalized each classifier's output values using the softmax function.

- *SVM:* A linear SVM was trained on the real-valued outputs of the component classifiers.
- *Feature Stacking:* The components were extracted from the image and their pixel values were combined into a single feature vector which was classified by a linear SVM.

The results for the four combination strategies are shown in the Table 2. Among the four combination techniques, the linear SVM performs slightly better than the maximum product and the maximum sum classifiers, the majority decision performed the worst. The table further shows the results for a system which used the face detection components in both the detection and identification stage ("SVM, linear, detection components").[11] The entry labeled "SVM, linear, single global component" refers to a system which used the component-based detector to localize the face and the previously described global component for identification. This system was therefor a combination of a component-based detector with a global identification classifier. The bottom three rows in the table show results for global systems, i.e. face identification systems which used the results of the global face detector to locate the face and extract the features. The second-degree polynomial SVM performed better than the LDA and the PCA. The best component systems improves on the best results of the global classifiers by 25%. Examples of correctly classified faces and misclassified faces are shown in Figs. 13 and 14. Table 3 shows the confusion matrix for the SVM component classifier. The error distribution among the ten subjects was highly unbalanced. The recognition rate for the tenth subject was as low as 35%. This might be explained by an inaccurate 3D head model or by the fact that this subject's training and test data were recorded six months apart from each other.

In our final experiment we tested the robustness of the systems against occlusions. We generated a new test set by pasting a rectangular patch of uniform gray value into each test image. The dimensions of each patch, its gay value and its location were chosen randomly for each test image.[12] Some example images of the test set with occlusions are shown in Fig. 15. The training set remained unchanged from the previous experiment.

The results for the occluded test images are shown in Table 4. The recognition rate of the component system dropped by around 20%, the recognition rate of the global system dropped by about 10%. We ran a second test in which the locations of the components were taken from the original test on occlusion-free images and fed to the identification module which ran on to the occluded test images. In this case the performance dropped by 10% to 80%, indicating that detection and identification stages contribute equally to the overall loss in performance. Although the component system still outperforms the global system by 15%, we expected a larger margin based on

Table 3. The confusion matrix for the component-based identification system with a linear SVM as combination classifier. The average recognition rate was 89.25%.

Subj. #	1	2	3	4	5	6	7	8	9	10
1	**183**	0	0	0	1	16	0	0	0	0
2	0	**199**	0	0	0	1	0	0	0	0
3	0	0	**200**	0	0	0	0	0	0	0
4	0	2	0	**197**	0	0	1	0	0	0
5	1	0	0	0	**196**	0	0	3	0	0
6	0	0	0	0	0	**200**	0	0	0	0
7	0	0	0	0	1	0	**199**	0	0	0
8	0	0	0	0	0	1	12	**181**	0	6
9	10	0	1	11	0	5	2	12	**159**	0
10	56	0	0	8	53	0	0	12	0	**71**

Figure 15. Examples from the test set with occlusions.

Table 4. The recognition rates on the occluded test set.

Classifier	Correct [%]	Errors [%]
Component, SVM, linear	69.0	31.0
Global, SVM, 2nd degree polynomial	54.65	45.35
Component, SVM, linear, No occlusions in the detection stage	79.65	20.35

our previous tests and the built-in advantage of local features over global features in images with occlusions. The steep drop in performance of the component-based system might be explained by the relatively large size of the components and their strong overlaps: even a small occlusion might affect several components simultaneously.

6. Conclusion

We described a component-based system for face detection and identification. The detection and identification modules shared the same two-layered architecture. In the first layer, component classifiers independently detected/identified parts of the face. The second layer contained a single combination classifier which combined the results of the component classifiers and performed the final detection/identification. We investigated several possibilities of including spatial information about the location of the components in the detection process. The best performance was achieved with a system in which the detection of the components was confined to small regions around the expected positions of the components. We also described a new method for learning relevant components for face detection and identification which is based on iteratively growing components in directions which minimize the error on a cross-validation set.

For face identification we used the component-based face detector in the training and testing stages to find the face in the image and to locate a set of reference points within the face. Around these points we extracted components specifically learned for identifying faces and classified them with our two-layered identification module. Two separate tests on a face detection and identification database showed that the component-based detector by itself and the combination of component-based detection and identification modules outperformed the global classifiers. In both cases we achieved improvements in the classification accuracy of about 25% on a test set without occlusions and 15% on a test set with occlusions.

Acknowledgment

The authors would like to thank P. Ho, B. Weyrauch and J. Huang for their work on various aspects of the component-based face identification systems. Thanks to V. Blanz and T. Vetter for providing the 3D head models.

Notes

1. Readers who are interested in large scale face identification databases are referred to Gross (2005) and Phillips et al. (2005).
2. The faces were rotated in depth from -30° and 30° rotation in depth in 5° steps. The faces were illuminated by ambient light and a single directional light pointing towards the center of the face. The position of the light source varied between -30° and 30° in azimuth and between 30° and 60° in elevation.
3. In the training set, the faces were rotated in depth from -30° to 30° in 6° increments. In the CV set, the rotation in depth ranged from -33° to 27° in 6° increments with $4\pm^\circ$ rotation in the image plane. For training, we rendered the 3D heads with 28 illumination models at each pose, for validation we used 18 slightly different models.
4. In our previous system (Heisele et al., 2001) we used a bound on the expected generalization error which was computed on the training set to control the growing direction.
5. The expected position of a component is the mean position of the component determined across all training images.
6. Each test image was scaled 10 times in a range from 58×58 to 90×90 pixels. The 58×58 classification window was slid across the scaled images in steps of one pixel.
7. The classifier is called haarclassifier__11_frontal__11_default.xml in OpenCV RC1.
8. Yet another way of computing the ROC curve was used in Viola and Jones (2004), where layers of the hierarchy were removed to generate points on the ROC curve.
9. The training set was recorded with an Olympus C-3040 digital still camera with 3.3 megapixels resolution. The test set was recorded with a Sony DFW-VL 500 video camera with at a resolution of 640×480 pixels.
10. We ran experiments with PCA dimensions ranging from 30 to 400. Above 300 there was no significant improvement in classification performance.
11. As for the previous experiments, a global component was added for face identification.
12. The patch dimensions were uniformly chosen from an interval of 10% to 30% of the image dimensions.

References

Beymer, D.J. 1993. Face recognition under varying pose. Center for Biological and Computational Learning, M.I.T., Cambridge, MA, A.I. Memo 1461.

Bileschi, S.M. and Heisele, B. 2002. Advances in component-based face detection. In *Proceedings of Pattern Recognition with Support Vector Machines, First International Workshop, SVM 2002*, Niagara Falls, pp. 135–143.

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

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

Dorko, G. and Schmid, C. 2003. Selection of scale invariant neighborhoods for object class recognition. In *International Conference on Computer Vision (ICCV)*, pp. 634–640.

Fergus, R., Perona, P., and Zisserman, A. 2003. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, pp. 264–271.

Gross, R. 2005. Face databases. In *Handbook of Face Recognition*, A. S. Li (ed.), Springer, New York.

Heisele, B., Ho, P., Wu, J., and Poggio, T. 2003. Face recognition: Component-based versus global approaches. *Computer Vision and Image Understanding (CVIU)* 91(1–2):6–21.

Heisele, B., Serre, T., Mukherjee, S., and Poggio. T. 2003. Hierarchical classification and feature reduction for fast face detection with support vector machines. *Pattern Recognition*, 36(9):2007–2017.

Heisele, B., Serre, T., Pontil, M., Vetter, T., and Poggio, T. 2001. Categorization by learning and combining object parts. In *Neural Information Processing Systems (NIPS)*, Vancouver.

Ivanov, Y., Heisele, B., and Serre, T. 2004. Using component features for face recognition. In *Proceedings of the 6th International Conference on Automatic Face and Gesture Recognition*, pp. 421–426.

Lienhart, R., Kuranov, A., and Pisarevsky, V. 2003. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *DAGM'03, 25th Pattern Recognition Symposium*, pp. 297–304.

Lowe, D.G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Martinez, A.M. 2002. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6):748–763.

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

Morgenstern, C. and Heisele, B. 2003. Component-based recognition of objects in an office environment. Center for Biological and Computational Learning, M.I.T., Cambridge, MA, A.I. Memo 232.

Nefian, A. and Hayes, M. 1999. An embedded HMM-based approach for face detection and recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 6, pp. 3553–3556.

Oren, M., Papageorgiou, C., Sinha, P., Osuna, E., and Poggio. T. 1997. Pedestrian detection using wavelet templates. In *IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, pp. 193–199.

Osuna, E. 1998. Support vector machines: Training and applications. Ph.D. dissertation, MIT, Department of Electrical Engineering and Computer Science, Cambridge, MA.

OpenCV Online Reference Manual. 2006. The Intel Corporation.

Pentland, A., Mogghadam, B., and Starner, T. 1994. View-based and modular eigenspaces for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Phillips, P.J. 1998. Matching pursuit filters applied to face identification. *IEEE Transactions on Image Processing*, 7(8):1150–1164.

Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., and Worek, W. 2005. Overview of the face recognition grand challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 947–954.

Poggio, T. and Edelman, S. 1990. A network that learns to recognize 3-D objects. *Nature*, 343:163–266.

Riesenhuber, M. and Poggio, T. 1999. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025.

Rowley, H.A., Baluja, S., and Kanade. T. 1998. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38.

Schneiderman, H. and Kanade, T. 2000. A statistical method for 3D object detection applied to faces and cars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 746–751.

Serre, T., Wolf, L., and Poggio, T. 2005. Object recognition with features inspired by visual cortex. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 994–1000.

Sim, T., Baker, S., and Bsat, M. 2003. The CMU pose, illumination, and expression database. *IEEE Transactions Pattern Analysis and Machine Intelligence (PAMI)* 25(12):1615–1618.

Sung, K.-K. 1996. Learning and example selection for object and pattern recognition. Ph.D. dissertation, MIT, Artificial Intelligence Laboratory and Center for Biological and Computational Learning, Cambridge, MA.

Ullman, S., Vidal-Naquet, M., and Sali, E. 2002. Visual features of intermdediate complexity and their use in classification. *Nature Neuroscience*, 5(7):682–687.

Vetter, T. 1998. Synthesis of novel views from a single face. *International Journal of Computer Vision*, 28(2):103–116.

Viola, P. and Jones, M.J. 2004. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

Weber, M., Welling, W., and Perona, P. 2000. Towards automatic dscovery of object categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

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

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