Component-based Face Detection

Bernd Heisele[†‡] Thomas Serre[†] Massimiliano Pontil[§] Tomaso Poggio[†]
[†]Center for Biological and Computational Learning, M.I.T., Cambridge, MA, USA
[‡]Honda R&D Americas, Inc., Boston, MA, USA
[§]Department of Information Engineering, University of Siena, Siena, Italy
{heisele, serre, tp} @ai.mit.edu pontil@dii.unisi.it

Abstract

We present a component-based, trainable system for detecting frontal and near-frontal views of faces in still gray images. The system consists of a two-level hierarchy of Support Vector Machine (SVM) classifiers. On the first level, component classifiers independently detect components of a face. On the second level, a single classifier checks if the geometrical configuration of the detected components in the image matches a geometrical model of a face. We propose a method for automatically learning components by using 3-D head models. This approach has the advantage that no manual interaction is required for choosing and extracting components. Experiments show that the component-based system is significantly more robust against rotations in depth than a comparable system trained on whole face patterns.

1. Introduction

Over the past ten years face detection has been thoroughly studied in computer vision research for mainly two reasons. First, face detection has a number of interesting applications: It can be part of a face recognition system, a surveillance system, or a video-based computer/machine interface. Second, faces form a class of visually similar objects which simplifies the generally difficult task of object detection.

In the following we give a brief overview of face detection techniques in still gray images. Since there are no color and motion cues available, face detection boils down to a pure pattern recognition task. A method for detecting faces in gray images by combining clustering techniques with neural networks is proposed in [15]. It generates face and non-face prototypes by clustering a set of training images. The distances between an input pattern and the prototypes are classified by a Multi-Layer Perceptron. In [8] frontal faces are detected by a polynomial SVM classifier. A system able to deal with rotations in the image plane was proposed by [10]. It consists of two neural networks, one for estimating the orientation of the face, and another for detecting the derotated faces. The recognition step was improved [11] by arbitrating between independently trained networks of identical structure. A naïve Bayesian approach was taken in [12]. The method determines the empirical probabilities of the occurrence of small rectangular intensity patterns within the face image. In [13] the system was expanded to deal with frontal and profile views of faces by adding a separate classifier trained on profile views. Another probabilistic approach which detects small parts of faces is proposed in [6]. Local feature extractors are used to detect the eyes, the corner of the mouth, and the tip of the nose. The geometrical configuration of these features is matched with a model configuration by conditional search. A related method using statistical models is published in [9]. Local features are extracted by applying multi-scale and multi-orientation filters to the input image. The responses of the filters on the training set are modeled as Gaussian distributions. Detecting components has also been applied to face recognition. In [18] local features are computed on the nodes of an elastic grid. Separate templates for the eyes, the nose, and the mouth are matched in [1, 2]. Finally, a component-based approach for people detection using SVMs was proposed in [7].

There are three basic ideas behind part- or component-based detection of objects. First, some object classes can be described well by a few characteristic object parts and their geometrical relation. Second, the patterns of some object parts might vary less under pose changes than the pattern belonging to the whole object. Third, a component-based approach might be more robust against partial occlusions than a global approach. The two main problems of a component-based approach are how to choose the set of discriminatory object parts and how to model their geometrical configuration. The above mentioned approaches either manually define a set of components and model their geometrical configuration or uniformly partition the image into

components and assume statistical independence between them.

We propose a technique for learning relevant components from 3-D head models. The technique starts with a set of small seed regions that are gradually grown by minimizing a bound on the expected error probability of an SVM. This approach has the advantage that no manual interaction is required for choosing and extracting components from the training set. Once the components have been determined, we train a system consisting of a two-level hierarchy of SVM classifiers. On the first level, component classifiers independently detect facial components. On the second level, a single classifier checks if the geometrical configuration of the detected components in the image matches a geometrical model of a face.

The outline of the paper is as follows: Section 2 gives a brief overview of SVM learning. In Section 3 we describe the component-based face detection system. A method for automatically extracting components from synthetic face images is presented in Section 4. Section 5 contains experimental results and a comparison between the global and component-based approaches. Section 6 concludes the paper.

2. Learning with Support Vector Machines

In this section we outline the basic theory of SVMs [16]. SVMs perform pattern recognition for two-class problems by determining the separating hyperplane[1] with maximum distance to the closest points of the training set. These points are called support vectors. If the data is not linearly separable in the input space, a non-linear transformation $\Phi(\cdot)$ can be applied which maps the data points $\mathbf{x} \in \mathbb{R}^n$ of the input space into a high (possibly infinite) dimensional space $\mathbb{R}^p$ which is called feature space. The data in the feature space is then separated by the optimal hyperplane as described above. The mapping $\Phi(\cdot)$ is implemented in the SVM classifier by a kernel function $K(\cdot,\cdot)$ which defines an inner product in $\mathbb{R}^N$, i.e. $K(\mathbf{x},\mathbf{t}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{t})$. The decision function of the SVM has the form:

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}), \tag{1}$$

where ℓ is the number of data points in the training set, and $y_i \in \{-1, 1\}$ is the class label of the data point $\mathbf{x}_i$. The coefficients α_i in Eq. (1) are the solution of a quadratic programming problem [16].

Let M be twice the distance of the support vectors to the hyperplane. This quantity is called margin and is given:

$$M = \frac{1}{\sqrt{\sum_{i=1}^{\ell} \alpha_i}}. \tag{2}$$

[1] SVM theory also includes the case of non-separable data, see [16].

The margin is an indicator of the separability of the data. In fact, the expected error probability of the SVM, EP_{err}, satisfies the following bound [16]:

$$EP_{err} \leq \frac{1}{\ell} E\left[\frac{D^2}{M^2}\right], \tag{3}$$

with D being the diameter of the smallest sphere containing the data points in the feature space. Later in the paper we will attempt to minimize this quantity to automatically extract components.

3. Component-based face detection

3.1. Motivation

We briefly mentioned in the introduction that a global approach is highly sensitive to changes in the pose of an object. Fig. 1 illustrates this problem for the simple case of linear classification. The result of training a linear classifier on faces can be represented as a single face template, schematically drawn in Fig. 1 a). Even for small rotations the template clearly deviates from the rotated faces as shown in Fig. 1 b) and c). The component-based approach tries to avoid this problem by independently detecting parts of the face. In Fig. 2 the eyes, nose, and the mouth are represented as single templates. For small rotations the changes in the components are small compared to the changes in whole face pattern. Slightly shifting the components is sufficient to achieve a reasonable match with the rotated faces.

Figure 1. Matching with a single template. The schematic template of a frontal face is shown in a). Slight rotations of the face in the image plane b) and in depth c) lead to considerable discrepancies between template and face.

3.2. Overview of the System

An overview of our two-level component-based classifier is shown in Fig. 3. On the first level, component classifiers independently detect components of the face. In the example shown these components are the eyes, the nose and the mouth. We used linear SVM classifiers, each of which was trained on a set of extracted facial components and on a set of randomly selected non-face patterns. The components were automatically extracted from synthetic 58×58

Figure 2. Matching with a set of component templates. The schematic component templates for a frontal face are shown in a). Shifting the component templates can compensate for slight rotations of the face in the image plane b) and in depth c).

face images generated from 3-D head models. On the second level the geometrical configuration classifier performs the final face detection by linearly combining the results of the component classifiers. Given a 58×58 window, the maximum continuous outputs of the component classifiers within rectangular search regions[2] around the expected positions of the components are used as inputs to the geometrical configuration classifier. The search regions have been calculated from the mean and standard deviation of the locations of the components in the training images. We also provide the geometrical classifier with the precise positions of the detected components relative to the upper left corner of the 58×58 window. Overall we have three values per component classifier that are propagated to the geometrical classifier. The system is computed as follows: We denote the input image as $\mathbf{x}$ and the extracted components as $\{\mathbf{x}^t\}_{t=1}^{T}$. The decision function of a component classifier is then given by:

$$f^t(\mathbf{x}^t) = \sum_{i=1}^{\ell} \alpha_i^t K^t(\mathbf{x}_i^t, \mathbf{x}^t).$$

where K^t is the kernel used by the t-th classifier. The geometrical configuration classifier $F(\mathbf{x})$ is a linear combination of the outputs of the component classifiers and the image locations (h^t, v^t) of the detected components:

$$F(\mathbf{x}) = \sum_{t=1}^{T} \mathbf{c}^t \cdot \left(f^t(\mathbf{x}^t), h^t, v^t\right)^T,$$

The coefficient vectors $\mathbf{c}^t$ are learned from the examples:

$$\left\{(f^1(\mathbf{x}_i^1), h_i^1, v_i^1, \ldots, f^T(\mathbf{x}_i^T), h_i^T, v_i^T, y_i)\right\}_{i=1}^{\ell}$$

where the label y_i is 1 for faces and -1 for non-face examples and ℓ is the number of examples.

[2]To account for changes in the size of the components, the outputs were determined over multiple scales of the input image. In our tests, we set the range of scales to $[0.75, 1.2]$.

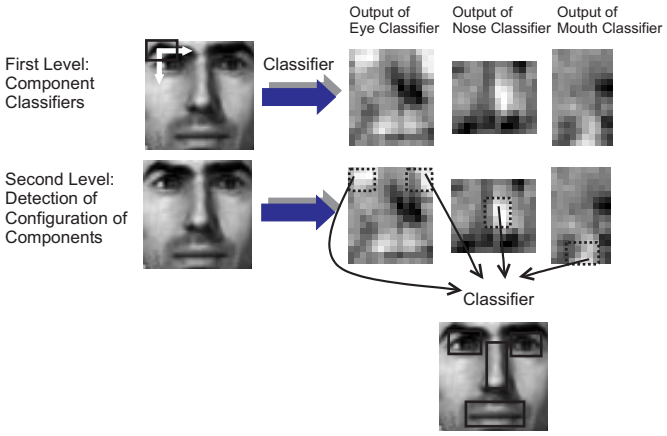

Figure 3. System overview of the component-based classifier using four components. On the first level, windows of the size of the components (solid lined boxes) are shifted over the face image and classified by the component classifiers. On the second level, the maximum outputs of the component classifiers within predefined search regions (dotted lined boxes) and the positions of the components are fed into the geometrical configuration classifier.

3.3. Training Data

Extracting face patterns is usually a tedious and time-consuming work that has to be done manually. Taking the component-based approach we would have to manually extract each single component from all images in the training set. This procedure would only be feasible for a small number of components. For this reason we used textured 3-D head models [17] to generate the training data. By rendering the 3-D head models we could automatically generate large numbers of faces in arbitrary poses and with arbitrary illumination. In addition to the 3-D information we also knew the 3-D correspondences for a set of reference points shown in Fig. 4. These correspondences allowed us to automatically extract facial components located around the reference points. Originally we had 7 textured head models acquired by a 3-D scanner. Additional head models were generated by 3-D morphing between all pairs of the original head models. The heads were rotated between -30° and 30° in depth. The faces were illuminated by ambient light and a single directional light pointing towards the center of the face, some examples are shown in Fig. 5. The position of the light varied between -30° and 30° in azimuth and between 30° and 60° in elevation. Overall, we generated 2,457 face images of size 58×58.

The negative training set initially consisted of 10,209 58×58 non-face patterns randomly extracted from 502 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 502 non-face images. The false positives (FPs) were added to the non-face training data to build the final non-face training set of size 13,654.

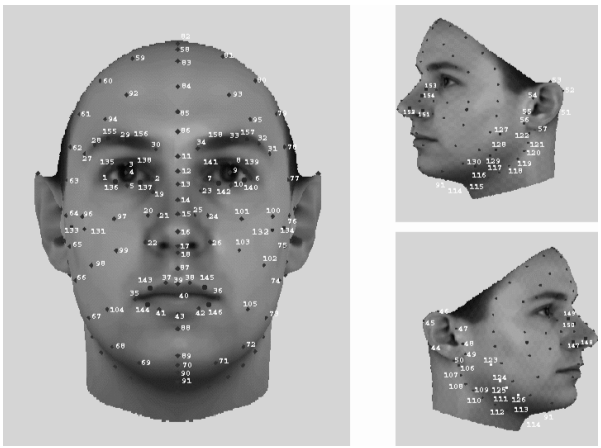

Figure 4. Reference points on the head models which were used for 3-D morphing and automatic extraction of facial components.

Figure 5. Examples of synthetic faces.

4. Learning Components

A main problem of the component-based approach is how to choose the set of discriminatory object parts. For the class of faces an obvious choice of components would be the eyes, the nose and the mouth. However, for other classes of objects it might be more difficult to manually define a set of intuitively meaningful components. Instead of manually choosing the components it would be more sensible to choose the components automatically based on their discriminative power and their robustness against pose and illumination changes.

Training a large number of classifiers on components of random size and location is one way to approach the problem of automatically determining components. The components can be ranked and selected based on the training results of the classifiers, e.g. the bound on the expected error probability. However, this method is computational extensive in the training stage.

An alternative to using a large set of arbitrary components is to specifically generate discriminative components. Following this idea, we developed a method that automatically determines rectangular components from a set of synthetic face images. The algorithm starts with a small rectangular component located around a pre-selected point in the face (e.g. center of the left eye). Note that we could locate the same facial point in all face images since we knew the point-by-point correspondences between the 3-D head models. The component is extracted from all synthetic face images to build a training set of positive examples. We also generate a training set of non-face patterns that have the same rectangular shape as the component. After training an SVM on the component data we estimate the performance of the SVM based on the estimated upper bound on the expected probability of error. According to Eq. (3) we calculate:

$$\rho = \frac{D^2}{M^2}, \tag{4}$$

where D is the diameter of the smallest sphere[3] in the feature space $\mathbb{R}^p$ containing the support vectors, and M is the margin given by Eq. (2). After determining ρ we enlarge the component by expanding the rectangle by one pixel into one of the four directions (up, down, left, right). Again, we generate training data, train an SVM and determine ρ. We do this for expansions into all four directions and finally keep the expansion which decreases ρ the most. This process is continued until the expansions into all four directions lead to an increase of ρ. In our experiments we started with 14 seed regions of size 5×5 most of them located in the vicinity of the eyes, nose and mouth. Fig. 6 shows the results after component growing; the size of the components is given in Table 4.

5. Experiments

[3] In our experiments we replaced D^2 in Eq. (4) by the dimensionality p of the feature space. This because our data points lay within an p-dimensional cube of length 1, so the smallest sphere containing the data had radius equal to $\sqrt{p}/2$. This approximation was mainly for computational reasons as in order to compute D we need to solve an optimization problem [8].

Components	Width	Height
Eyebrows	19	15
Eyes	17	17
Between eyes	18	16
Nose	15	20
Nostrils	22	12
Cheeks	21	20
Mouth	31	15
Lip	13	16
Corners of the mouth	18	11

Table 1. Size of the learned components.

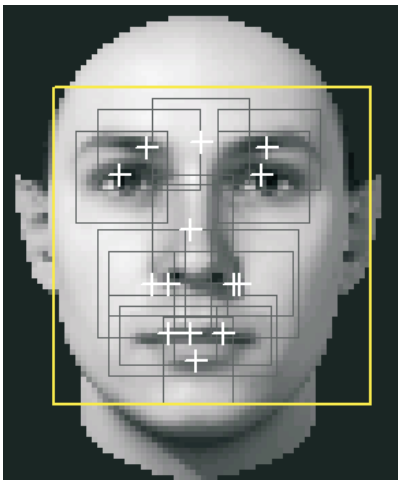

Figure 6. The fourteen learned components. The crosses mark the centers of the components.

In our experiments we compared the component-based system to a classifier trained on the whole face pattern. The component system consisted of 14 linear SVM classifiers for component detection and a single linear SVM as geometrical classifier. The whole face classifier was a single linear SVM trained on gray values of the whole face pattern. The training data for both classifiers consisted of 2,457 synthetic gray face images and 13,655 non-face gray images of size 58×58.

The positive test consisted of 1,834 faces rotated between about -30° and 30° in depth. The faces were manually extracted from the CMU PIE database [14]. The negative test set consisted of 24,464 difficult non-face patterns that were collected by a fast face detector [5] from web images[4]. The false positive (FP) rate was calculated relative to the number of non-face test images. The comparison between SVM whole face classifiers (linear and polynomial kernels) and a component classifier consisting of 14 linear SVM component classifiers and a linear SVM geometrical configuration classifier is shown in Fig. 7. For benchmarking we also added the ROC curve of a second-degree polynomial kernel SVM trained on 19×19 real face images. This face detector is described and evaluated in detail in [3] and performed amongst the best face detection systems on the CMU test set [10] including frontal and near-frontal face images. The component system outperforms all whole face systems. Some detection results generated by the component system are shown in Fig. 8.

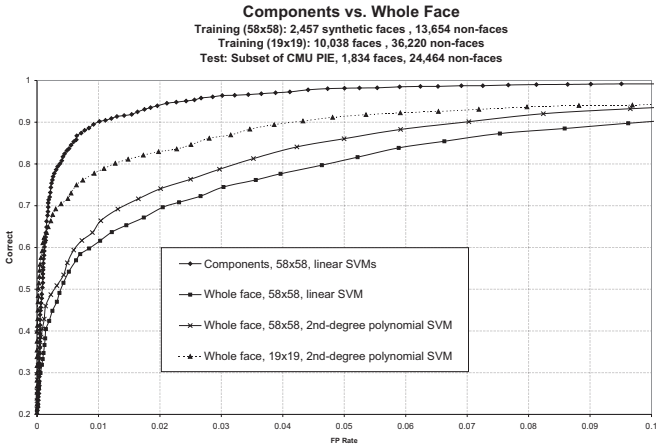

Figure 7. ROC curves for whole face classifiers and the 14 component classifier.

6. Conclusion

We presented a component-based system for face detection using SVM classifiers. The system performs the detection by means of a two level hierarchy of classifiers. On the first level, the component classifiers independently detect parts of the face. On the second level, the geometrical configuration classifier combines the results of the component classifiers and performs the final detection step. Experiments on real face images show a significant improvement in the classification performance compared to a whole face detection system. We also proposed a region growing method that involves measures derived from SVM theory to learn relevant components from a set of 3-D head models. The use of 3-D head models allowed us to automatically extract components and to arbitrarily change the illumination and the viewpoint. Both, the component-based classification system and the technique for component learning can be applied to other object detection tasks in computer vision.

[4]The test database together with a detailed description of the experiments [4] can be found on the MIT/CBCL web page.

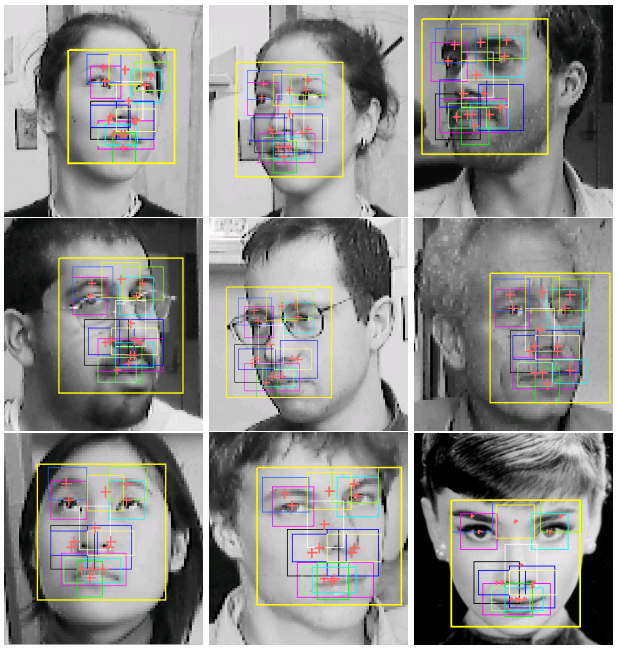

Figure 8. Faces detected by the 14 component system.

Acknowledgements

The authors would like to thank V. Blanz and T. Vetter for providing the 3-D models. The research was partially sponsored by DARPA under contract No. N00014-00-1-0907 and the National Science Foundation under contract No. IIS-9800032. Additional support was provided by the DFG, Eastman Kodak, Compaq, and Honda R&D.

References

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

[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, T. Poggio, and M. Pontil. Face detection in still gray images. AI Memo 1687, Center for Biological and Computational Learning, MIT, Cambridge, MA, 2000.

[4] B. Heisele, T. Serre, S. Mukherjee, and T. Poggio. Feature reduction and hierarchy of classifiers for fast object detection in video images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii, 2001.

[5] B. Heisele, T. Serre, and T. Poggio. Component-based face detection. AI Memo, Center for Biological and Computational Learning, MIT, Cambridge, MA, 2001.

[6] T. K. Leung, M. C. Burl, and P. Perona. Finding faces in cluttered scenes using random labeled graph matching. In *Proc. International Conference on Computer Vision*, pages 637–644, Cambridge, MA, 1995.

[7] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4):349–361, 2001.

[8] E. Osuna. *Support Vector Machines: Training and Applications*. PhD thesis, MIT, EE/CS Dept., Cambridge, MA, 1998.

[9] T. D. Rikert, M. J. Jones, and P. Viola. A cluster-based statistical model for object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1046–1053, Fort Collins, 1999.

[10] H. A. Rowley, S. Baluja, and T. Kanade. Rotation invariant neural network-based face detection. Computer Scienct Technical Report CMU-CS-97-201, CMU, Pittsburgh, 1997.

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

[12] H. Schneiderman and T. Kanade. Probabilistic modeling of local appearance and spatial relationships for object recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 45–51, Santa Barbara, 1998.

[13] 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.

[14] T. Sim, S. Baker, and M. Bsat. The cmu Pose, Illumination, and Expression (PIE) database of human faces. Computer Science Technical Report 01-02, CMU, Pittsburgh, 2001.

[15] K.-K. Sung. *Learning and Example Selection for Object and Pattern Recognition*. PhD thesis, MIT, AI Lab, Cambridge, MA, 1996.

[16] V. Vapnik. *Statistical learning theory*. John Wiley and Sons, New York, 1998.

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

[18] L. Wiskott, J. M. Fellous, N. Kruger, and C. von der Malsburg. A statistical method for 3d object detection applied to faces and cars. In *Proc. IEEE International Conference on Image Processing*, pages 129–132, 1997.