Categorization by Learning and Combining Object Parts

Bernd Heisele [†] **Thomas Serre** **Massimiliano Pontil**[‡] **Thomas Vetter**[§] **Tomaso Poggio**
Center for Biological and Computational Learning, M.I.T., USA
45 Carleton St, Cambridge, MA 02142
{*heisele,serre*}*@ai.mit.edu* *pontil@ing.unisi.it* *vetter@informatik.uni-freiburg.de* *tp@ai.mit.edu*

Abstract

We describe an algorithm for automatically learning discriminative parts in object images with SVM classifiers. It is based on growing image parts by minimizing theoretical bounds on the error probability of an SVM. Component-based face classifiers are then combined in a second stage to yield a hierarchical SVM classifier. Experimental results in face classification show considerable robustness for rotations in depth and suggest performance at significantly better level than other face detection systems. Novel aspects of our approach are: a) an algorithm to learn from examples component-based classification experts and their combination, b) the use of 3D morphable models for training and c) a MAX operation – on the output of each component classifier within a search region – which may be relevant for biological models of visual recognition.

1 Introduction

We study the problem of automatically synthesizing hierarchical classifiers by learning discriminative object parts in images. Our motivation is that most object classes (e.g. faces, horses, cars) seem to be naturally described by a few characteristic parts or components and their geometrical relation. Greater invariance to viewpoint changes and robustness against partial occlusions is the two main potential advantages of component-based approaches.

The first challenge in developing component-based systems is how to choose automatically a set of discriminatory object parts. Instead of manually selecting the components, it is desirable to learn the components from a set of examples based on their discriminative power and their robustness against pose and illumination changes. The second challenge is to combine the component-based experts by learning implicitly their geometrical configuration.

[†]Honda Research Laboratory, Boston, MA, from April 2001
[‡]Department of Information Engineering, University of Siena, Italy
[§]Computer Graphics Research Group, University of Freiburg, Germany

2 Background

Global approach by detecting object as a single template was successfully applied to tasks where the pose of the object was fixed. In [5] Haar wavelet features were used to detect frontal and back views of pedestrians with a SVM classifier. Learning-based system for detecting frontal faces based on a gray value features are described in [14, 12, 9, 2].

Component-based techniques promise to provide more invariance since the individual components vary less while the variations induced by pose changes occur mainly in their geometry. A component-based method for detecting faces based on the empirical probabilities of overlapping rectangular parts of the image was proposed in [10]. Another probabilistic approach, which detects small parts of faces, is proposed in [3] by using local feature extractors to detect the eyes, the corner of the mouth and 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 [8]. 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. In [4] pedestrian detection was performed by a set of SVM classifiers each of which was trained to detect a specific part of the human body. Other approaches similar to ours select components on the basis of information-based criteria [13, 17].

In this paper we present a technique for learning relevant object components. It starts with a set of small seed regions that are gradually grown by minimizing a bound on the expected error probability of an SVM. Once the components have been determined, we train a system consisting of a two-level hierarchy of SVM classifiers. First, component classifiers independently detect facial components. Second, a combination classifier learns implicitly the geometry of the components.

3 Learning Components with Support Vector Machines

3.1 Linear Support Vector Machines[1]

Linear SVMs [15] perform pattern recognition for two-class problems by determining the separating hyperplane with maximum distance to the closest points of the training set. These points are called support vectors. The decision function of the SVM has the form:

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

where ℓ is the number of support vectors, and $y_i \in \{-1, 1\}$ is the class label of the training point, α_i some positive parameters and $\mathbf{x} \in \mathcal{R}^p$. They are the solution of a quadratic programming problem. Let M be twice the distance of the support vectors to the hyperplane. This quantity is called margin and is given by:

$$M = \frac{2}{\sqrt{\sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j < \mathbf{x}_i \cdot \mathbf{x}_j >}}. \tag{2}$$

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 [15]:

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

where D is the diameter of the smallest sphere containing the data points in space $\mathcal{R}^p$ and the expectation E is taken with respect to the probability of the training set.

[1] For simplicity we deal with linear SVMs (in the separable case) only but the algorithm discuss below can be applied to non-linear SVMs as well.

3.2 Learning Components

We describe a method that automatically determines rectangular components from a set of object images. The algorithm starts with a small rectangular component located around a pre-selected (possibly at random) point in the object image (e.g., in the case of face images, this could be the center of the left eye). The component is extracted from each object images to build a training set of positive examples[1]. 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 upper bound on the error probability. According to Eq. (3) we estimate:

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

As shown in [15], this quantity can be computed by solving a quadratic programming problem. After determining ρ we grow 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 the SVM and compute ρ. We do this for expansions into each of the four directions and continue the process until ρ increases. The same greedy process can then be repeated by selecting each time a new seed region. The set of extracted components is then ranked according to the final value of $\frac{D}{M}$ and the top N are then chosen. Bounds such as the one in Eq. (3) have been recently used for feature selection [18].

4 Learning Facial Components

We applied the above algorithm for extracting and selecting components from a training set of face images. We used 7 textured 3-D head models [16] acquired by a 3-D scanner and set in voxel-wise correspondence, to obtain 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. We also knew the 3-D correspondences for a set of reference points shown in Fig. 1a. These correspondences allowed us to automatically extract facial components located around the reference points in the images. Additional head models were generated by 3-D morphing between all pairs of the original 7 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. 1b. 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 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 difficult to classify [12]. 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. All images were normalized in the $[0, 1]$ interval.

We started with 100 random seed regions of size 5×5^2. Most of the top-ranked components were found in the vicinity of the eyes, nose and mouth.

[1] We suppose that the point-by-point correspondences between object images are given. We have this information for our training set (see later). It is also possible to bootstrap the point-by-point correspondence from just a few images already in correspondence to the rest of the training set.

[2] In the experiments we replaced D^2 in Eq. (4) by the dimensionality p of space $\mathcal{R}^p$, since our data points lay within an p-dimensional cube of length 1 and thus the smallest sphere containing the data had radius equal to $\sqrt{p}/2$. This approximation was motivated by computational reasons.

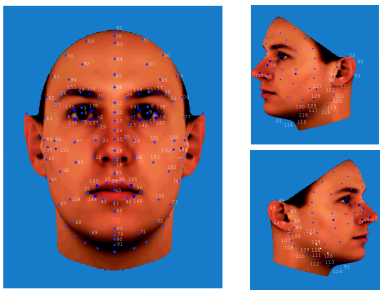
a)  b) 

Figure 1: a) Reference points on the head models, which were used for 3-D morphing and automatic extraction of facial components. b) Examples of synthetic faces.

5 Combining Components

An overview of our two-level component-based classifier is shown in Fig. 2. On the first level, component classifiers independently detect components of the face. only 4 components out of 14) these components are the eyes, the nose and the mouth. Each classifier was trained on a set of facial components and on a set of randomly selected non-face patterns. On the second level the geometrical configuration classifier performs face detection by combining the component classifiers. The maximum real-valued outputs of each component classifier within its rectangular search region[3] around the expected positions of the components are used as inputs to the combination classifier. The size of the search regions was estimated from the mean and the standard deviation of the locations of the components in the training images. The MAX operation is performed done both during training and at run time. Interestingly it turns out to be similar to the key pooling mechanism postulated in a recent model of object recognition in visual cortex [7].

The information flow is 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: $f^t(\mathbf{x}^t) = \sum_{i=1}^{\ell} \alpha_i^t y_i^t < \mathbf{x}_i^t \cdot \mathbf{x}^t >$.

Then the maximum of each component classifier within its search region is selected and its value together with the associated image locations (h^t, v^t) is fed to the combination classifier – a linear SVM. Thus $F_c(\mathbf{x}) = \sum_{t=1}^{T} \mathbf{c_1}^t \cdot (f^t(\mathbf{x}^t), h^t, v^t)^T$. The coefficient vectors $\mathbf{c_1}^t$ are learned from the examples: $\{(f^1(\mathbf{x}_i^1), h_i^1, v_i^1, \ldots, f^T(\mathbf{x}_i^T), h_i^T, v_i^T); y_i\}_{i=1}^{\ell}$.

6 Experiments

The component "experts" consisted of 14 linear SVM classifiers for component detection and a single linear SVM as geometrical classifier (see Section 5). For reference, we trained a whole face classifier (single linear SVM) on gray values of the whole face pattern. The training data for all classifiers consisted of 2,457 synthetic gray face images and 13,654 non-face gray images of size 58×58 (see 4. The positive test set consisted of 1,834 gray images of real faces from the 41,368 images of faces + background of the new CMU PIE test set[11]. We manually extracted images suitable for testing a face classifiers: some examples are shown in figure 3. We plan to make this database available on our web site http://www.ai.mit.edu/projects/cbcl/software-datasets/index.html. The database includes faces rotated between about $-30°$ and $30°$. The negative test set consists of

[3]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 $[0.75, 1.2]$.

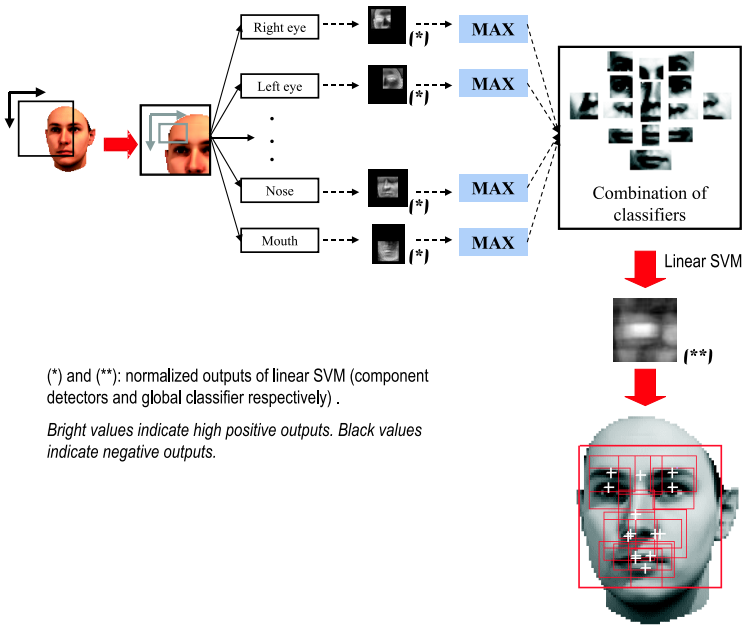


Figure 2: System overview of the component-based classifier. At the first level, each of component classifiers is run at multiple positions within its search region. The maximum of its outputs is selected and fed, together with the associated position, to the combination classifier.

24,464 non-face patterns selected by a combination of low-resolution SVM classifiers as the most similar to non-faces from background images collected on the web. The test was performed in the following way: An image of the face test set was scaled down 13 times to be in a range from 58×58 to 105×105 pixels. Images of the non-face test set was all processed at scale 1.0. After rescaling the images, a window of size 58×58 was shifted pixel-by-pixel over each of the scaled images. As proposed in [12] we applied two preprocessing steps to the 8 bit gray values within the 58×58 window. First, a best-fit intensity plane was subtracted from the gray values to compensate for cast shadows. Then histogram equalization was applied to remove variations in the image brightness and contrast[4]. The resulting pixel values were scaled to be in a range between 0 and 1. We generated the ROC curve by considering only the maximum outputs for each patterns (faces and non-faces) as the detection result for every scales and positions within the pattern. The comparison between a 58×58 linear SVM and the new component-based system is shown in Fig. 4a.

A natural question that arises is about the role of geometrical information. To begin answer this question – which has relevant implications for models of cortex – we tested another system in which the combination classifier receives as inputs only the output of each component classifier but not the position of its maximum (within the search region). In this case $F'_c(\mathbf{x}) = \sum_{t=1}^{T} \mathbf{c_2}^t \cdot (f^t(\mathbf{x}^t))^T$. The coefficient vectors $\mathbf{c_2}^t$ are learned from the examples: $\{(f^1(\mathbf{x}_i^1), \ldots, f^T(\mathbf{x}_i^T); y_i\}_{i=1}^{\ell}$, where the label y_i is 1 for faces and -1 for non-face examples and ℓ is the number of examples. This system, with more limited position information,

[4]For components the histogram equalization was performed on a 58×58 which was positioned such that the component window was at the expected position of the component.

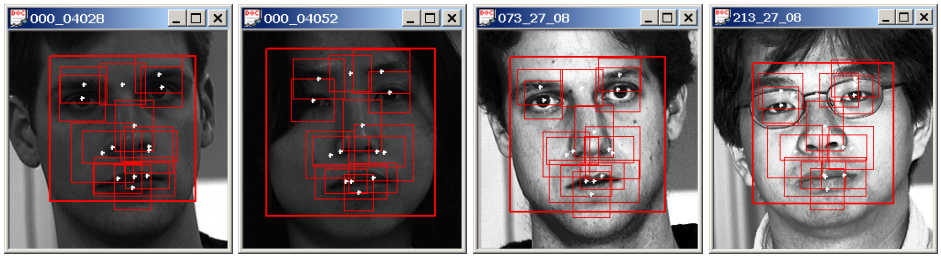
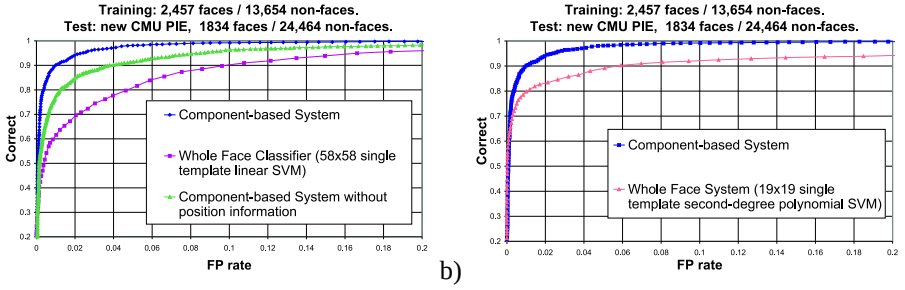


Figure 3: Faces from the new CMU PIE test set detected by the 14 component-based classifier.

also outperforms the whole face system but is significantly inferior in performance to our basic hierarchical classifier that uses position information.

The classification results show that the component-based system is significantly more robust against rotations in depth than comparable systems trained on whole face patterns. Because of the resolution required by the component-based system, a direct comparison with other published systems on the standard MIT-CMU test set [9] is impossible. For an indirect comparison, we used the 19×19 2^{nd} degree SVM whole face classifier [2] of which we know the performance on both the MIT-CMU test set and also on the new CMU PIE test set. At 90% correct (see Fig. 4b) that system has 75 FPs (see [2]) on the CMU-MIT test set; the same system has about 5 times more FPs than our best component-based system on the new CMU PIE test set. If the relative performances between the two systems remains the same on the MIT-CMU test set, we would expect about 15 FPs on the MIT-CMU database for our component-based system. This number should be compared with the 33 FPs at 90.5% correct of [10] which is so far the best face detection system.



a) b)

Figure 4: a) ROC curves for a 58×58 single template linear SVM classifier and the component-based system. For comparison a component-based classifier trained without use of direct position information has been added. The FPs are given relative to the number of total patterns. b) Comparison with a reference 19×19 2^{nd} degree polynomial SVM classifier and the component-based system.

7 Open Questions

An extension under way of the component-based approach to face identification is already showing good performances [1]. Another natural generalization of the work described here involves the application of our system to various classes of objects such as cars, animals and people. Still another extension regards the question of view-invariant object detection. As suggested by [6] in a biological context and demonstrated recently by [10] full pose

invariance in recognition tasks can be achieved by combining view-dependent classifiers. It is interesting to ask whether the approach described here could also be used to learn which views are most discriminative and how to combine them optimally. Finally, the role of geometry, and in particular how to compute and represent position information in biologically plausible networks is an important open question at the interface between machine and biological vision.

References

[1] B. Heisele, P. Ho, and T. Poggio. Face recognition with support vector machines: Global versus component-based approach. In *ICCV*, 2001.

[2] B. Heisele, T. Poggio, and M. Pontil. Face detection in still gray images. A.I. memo 1687, Center for Biological and Computational Learning, MIT, Cambridge, MA, 2000.

[3] 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.

[4] 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.

[5] C. Papageorgiou and T. Poggio. A trainable system for object detection. In *International Journal of Computer Vision*, volume 38, 1, pages 15–33, 2000.

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

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

[8] 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.

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

[10] 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.

[11] 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, 2001.

[12] K.-K. Sung. *Learning and Example Selection for Object and Pattern Recognition*. PhD thesis, MIT, Artificial Intelligence Laboratory and Center for Biological and Computational Learning, Cambridge, MA, 1996.

[13] S. Ullman and E. Sali. Object classification using a fragment-based representation. In *Biologically Motivated Computer Vision (eds. S.-W. Lee, H. Bulthoff ad T. Poggio)*, pages 73–87 (Springer, New York), 2000.

[14] R. Vaillant, C. Monrocq, and Y. Le Cun. An original approach for the localisation of objects in images. In *International Conference on Artificial Neural Networks*, pages 26–30, 1993.

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

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

[17] M. Weber, W. Welling, and P. Perona. Towards automatic dscovery of object categories. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2000.

[18] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for support vector machines. In *Advances in Neural Information Processing Systems 13*, 2001.