Face Recognition with Support Vector Machines: Global versus Component-based Approach

Bernd Heisèle\textsuperscript{1} Purdy Ho\textsuperscript{1} Tomaso Poggio
Massachusetts Institute of Technology
Center for Biological and Computational Learning
Cambridge, MA 02142
heisèle@ai.mit.edu purdyho@mit.edu tp@ai.mit.edu

Abstract

We present a component-based method and two global methods for face recognition and evaluate them with respect to robustness against pose changes. In the component system we first locate facial components, extract them and combine them into a single feature vector which is classified by a Support Vector Machine (SVM). The two global systems recognize faces by classifying a single feature vector consisting of the gray values of the whole face image. In the first global system we trained a single SVM classifier for each person in the database. The second system consists of sets of viewpoint-specific SVM classifiers and involves clustering during training. We performed extensive tests on a database which included faces rotated up to about $45\degree$ in depth. The component system clearly outperformed both global systems on all tests.

1. Introduction

Over the past 20 years numerous face recognition papers have been published in the computer vision community; a survey can be found in [4]. The number of real-world applications (e.g. surveillance, secure access, human/computer interface) and the availability of cheap and powerful hardware also lead to the development of commercial face recognition systems. Despite the success of some of these systems in constrained scenarios, the general task of face recognition still poses a number of challenges with respect to changes in illumination, facial expression, and pose.

In the following we give a brief overview on face recognition methods. Focusing on the aspect of pose invariance we divide face recognition techniques into two categories: i) global approach and ii) component-based approach.

i) In this category a single feature vector that represents the whole face image is used as input to a classifier. Several classifiers have been proposed in the literature e.g. minimum distance classification in the eigenspace [18, 20], Fisher’s discriminant analysis [1], and neural networks [6]. Global techniques work well for classifying frontal views of faces. However, they are not robust against pose changes since global features are highly sensitive to translation and rotation of the face. To avoid this problem an alignment stage can be added before classifying the face. Aligning an input face image with a reference face image requires computing correspondences between the two face images. The correspondences are usually determined for a small number of prominent points in the face like the center of the eye, the nostrils, or the corners of the mouth. Based on these correspondences the input face image can be warped to a reference face image. In [12] an affine transformation is computed to perform the warping. Active shape models are used in [10] to align input faces with model faces. A semi-automatic alignment step in combination with SVM classification was proposed in [9].

ii) An alternative to the global approaches is to classify local facial components. The main idea of component-based recognition is to compensate for pose changes by allowing a flexible geometrical relation between the components in the classification stage. In [3] face recognition was performed by independently matching templates of three facial regions (both eyes, nose and 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 [2]. In [23] a geometrical model of a face was implemented by a 2-D elastic graph. The recognition was based on wavelet coefficients that were computed on the nodes of the elastic graph. In [14] a window was shifted

\textsuperscript{1}with Honda Research Laboratory, Boston, MA, from April 2001
\textsuperscript{2}with Hewlett-Packard, Palo Alto, CA, from July 2001
over the face image and the DCT coefficients computed within the window were fed into a 2-D Hidden Markov Model.

We present two global approaches and a component-based approach to face recognition and evaluate their robustness against pose changes. The first global method consists of a face detector which extracts the face part from an image and propagates it to a set of SVM classifiers that perform the face recognition. By using a face detector we achieve translation and scale invariance. In the second global method we split the images of each person into viewpoint-specific clusters. We then train SVM classifiers on each single cluster. In contrast to the global methods, the component system uses a face detector that detects and extracts local components of the face. The detector consists of a set of SVM classifiers that locate facial components and a single geometrical classifier that checks if the configuration of the components matches a learned geometrical face model. The detected components are extracted from the image, normalized in size and fed into a set of SVM classifiers.

The outline of the paper is as follows: Chapter 2 gives a brief overview on SVM learning and on strategies for multi-class classification with SVMs. In Chapter 3 we describe the two global methods for face recognition. Chapter 4 is about the component-based system. Chapter 5 contains experimental results and a comparison between the global and component systems. Chapter 6 concludes the paper.

2. Support Vector Machine Classification

We first explain the basics of SVMs for binary classification [21]. Then we discuss how this technique can be extended to deal with general multi-class classification problems.

2.1. Binary Classification

SVMs belong to the class of maximum margin classifiers. They perform pattern recognition between two classes by finding a decision surface that has maximum distance to the closest points in the training set which are termed support vectors. We start with a training set of points \(\mathbf{x}_i \in \mathbb{R}^n, i = 1, 2, \ldots, N\) where each point \(\mathbf{x}_i\) belongs to one of two classes identified by the label \(y_i \in \{-1, 1\}\). Assuming linearly separable data\(^1\), the goal of maximum margin classification is to separate the two classes by a hyperplane such that the distance to the support vectors is maximized. This hyperplane is called the optimal separating hyperplane (OSH). The OSH has the form:

\[
f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b, \tag{1}
\]

\(^1\)For the non-separable case the reader is referred to [21].

The coefficients \(\alpha_i\) and the \(b\) in Eq. (1) are the solutions of a quadratic programming problem [21]. Classification of a new data point \(\mathbf{x}\) is performed by computing the sign of the right side of Eq. (1). In the following we will use

\[
d(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i y_i x_i \cdot x + b \over \left\| \sum_{i=1}^{\ell} \alpha_i y_i x_i \right\| \tag{2}
\]

to perform multi-class classification. The sign of \(d\) is the classification result for \(\mathbf{x}\), and \(|d|\) is the distance from \(\mathbf{x}\) to the hyperplane. Intuitively, the farther away a point is from the decision surface, i.e. the larger \(|d|\), the more reliable the classification result.

The entire construction can be extended to the case of nonlinear separating surfaces. Each point \(\mathbf{x}\) in the input space is mapped to a point \(z = \phi(\mathbf{x})\) of a higher dimensional space, called the feature space, where the data are separated by a hyperplane. The key property in this construction is that the mapping \(\Phi(\cdot)\) is subject to the condition that the dot product of two points in the feature space \(\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})\) can be rewritten as a kernel function \(K(\mathbf{x}, \mathbf{y})\). The decision surface has the equation:

\[
f(\mathbf{x}) = \sum_{i=1}^{\ell} y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b, \tag{2}
\]

again, the coefficients \(\alpha_i\) and \(b\) are the solutions of a quadratic programming problem. Note that \(f(\mathbf{x})\) does not depend on the dimensionality of the feature space.

An important family of kernel functions is the polynomial kernel:

\[
K(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x} \cdot \mathbf{y})^d,
\]

where \(d\) is the degree of the polynomial. In this case the components of the mapping \(\Phi(\mathbf{x})\) are all the possible monomials of input components up to the degree \(d\).

2.2. Multi-class classification

There are two basic strategies for solving \(q\)-class problems with SVMs:

i) In the one-vs-all approach \(q\) SVMs are trained. Each of the SVMs separates a single class from all remaining classes [5, 17].

ii) In the pairwise approach \(\binom{q}{2}\) machines are trained. Each SVM separates a pair of classes. The pairwise classifiers are arranged in trees, where each tree node represents an SVM. A bottom-up tree similar to the elimination tree used in tennis tournaments was originally proposed in [16] for recognition of 3-D objects and was applied to face recognition in [7]. A top-down tree structure has been recently published in [15].

There is no theoretical analysis of the two strategies with respect to classification performance. Regarding the training effort, the one-vs-all approach is preferable since only \(q\)
SVMs have to be trained compared to $q^2$ SVMs in the pairwise approach. At run-time both strategies require the evaluation of $q-1$ SVMs. Recent experiments on person recognition show similar classification performances for the two strategies [13]. Since the number of classes in face recognition can be rather large we opted for the one-vs-all strategy where the number of SVMs is linear with the number of classes.

3. Global Approach

Both global systems described in this paper consist of a face detection stage where the face is detected and extracted from an input image and a recognition stage where the person’s identity is established.

3.1. Face detection

We developed a face detector similar to the one described in [8]. In order to detect faces at different scales we first computed a resolution pyramid for the input image and then shifted a $58 \times 58$ window over each image in the pyramid. We applied two preprocessing steps to the gray images to compensate for certain sources of image variations [19]. 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. The resulting gray values were normalized to be in a range between 0 and 1 and were used as input features to a linear SVM classifier. Some detection results are shown in Fig. 1.

The training data for the face detector were generated by rendering seven textured 3-D head models [22]. The heads were shown in Fig. 1. The training data for the face detector were generated by rendering seven textured 3-D head models [22]. The heads were shown in Fig. 1. The training data for the face detector were generated by rendering seven textured 3-D head models [22]. The heads were shown in Fig. 1. The training data for the face detector were generated by rendering seven textured 3-D head models [22]. The heads were shown in Fig. 1. The training data for the face detector were generated by rendering seven textured 3-D head models [22]. The heads were shown in Fig. 1. The training data for the face detector were generated by rendering seven textured 3-D head models [22]. The heads were shown in Fig. 1. The training data for the face detector were generated by rendering seven textured 3-D head models [22]. The heads were shown in Fig. 1. The training data for the face detector were generated by rendering seven textured 3-D head models [22]. The heads were shown in Fig. 1. The training data for the face detector were generated by rendering seven textured 3-D head models [22]. The heads were shown in Fig. 1. The training data for the face detector were generated by rendering seven textured 3-D head models [22]. The heads were shown in Fig. 1. The training data for the face detector were generated by rendering seven textured 3-D head models [22]. The heads were shown in Fig. 1. The training data for the face detector were generated by rendering seven textured 3-D head models [22]. The heads were shown in Fig. 1. The training data for the face detector were generated by rendering seven textured 3-D head models [22]. The heads were shown in Fig. 1. The training data for the face detector were generated by rendering seven textured 3-D head models [22]. The heads were shown in Fig. 1. The training data for the face detector were generated by rendering seven textured 3-D head models [22]. The heads were shown in Fig. 1. The training data for the face detector were generated by rendering seven textured 3-D head models [22]. The heads were shown in Fig. 1. The training data for the face detector were generated by rendering seven textured 3-D head models [22]. The heads were shown in Fig. 1. The training data for the face detector were generated by rendering seven textured 3-D head models [22]. The heads were shown in Fig. 1. The training data for the face detector were generated by rendering seven textured 3-D head modes...
distance to all other faces in the same cluster is chosen to be the average face of the cluster. Iterative clustering stops when a maximum number of clusters is reached. The average faces can be arranged in a binary tree. Fig. 2 shows the result of clustering applied to the training images of a person in our database. The nodes represent the average faces; the leaves of the tree are some example faces of the final clusters. As expected divisive clustering performs a viewpoint-specific grouping of faces.

We trained a linear SVM to distinguish between all images in one cluster (labeled +1) and all images of other people in the training set (labeled −1). Classification was done according to Eq. (3) with q now being the number of clusters of all people in the training set.

4. Component-based Approach

The global approach is highly sensitive to image variations caused by changes in the pose of the face. The component-based approach avoids this problem by independently detecting parts of the face. For small rotations, the changes in the components are relatively small compared to the changes in the whole face pattern. Changes in the 2-D locations of the components due to pose changes are accounted for by a learned, flexible face model.

4.1. Detection

We implemented a two-level component-based face detector which is described in detail in [8]. The principles of the system are illustrated in Fig. 3. On the first level, component classifiers independently detected facial components. On the second level, a geometrical configuration classifier performed the final face detection by combining the results of the component classifiers. Given a 58 × 58 window, the maximum continuous outputs of the component classifiers within rectangular search regions around the expected positions of the components were used as inputs to the geometrical configuration classifier. The search regions have been calculated from the mean and standard deviation of the components’ locations in the training images. We also provided the geometrical classifier with the precise positions of the detected components relative to the upper left corner of the 58 × 58 window. The 14 facial components used in the detection system are shown in Fig. 4 (a). The shapes and positions of the components have been automatically determined from the training data in order to provide maximum discrimination between face and non-face images; see [8] for details about the algorithm. The training set was the same as for the whole face detector described in the previous Chapter.
4.2. Recognition

To train the face recognizer we first ran the component-based detector over each image in the training set and extracted the components. From the 14 original we kept 10 for face recognition, removing those that either contained few gray value structures (e.g. cheeks) or strongly overlapped with other components. The 10 selected components are shown in Fig. 4 (b). Examples of the component-based face detector applied to images of the training set are shown in Fig. 5. To generate the input to our face recognition classifier we normalized each of the components in size and combined their gray values into a single feature vector5. As for the first global system we used a one-vs-all approach with a linear SVM for every person in the database. The classification result was determined according to Eq. (3).

5. Experiments

The training data for the face recognition system were recorded with a digital video camera at a frame rate of about 5 Hz. The training set consisted of 8,593 gray face images of five subjects from which 1,383 were frontal views. The resolution of the face images ranged between $80 \times 80$ and $130 \times 130$ pixels with rotations in azimuth up to about $\pm 40^\circ$. The test set was recorded with the same camera but on a separate day and under different illumination and with different background. The set included 974 images of all five subjects in our database. The rotations in depth was again up to about $\pm 40^\circ$.

Two experiments were carried out. In the first experiment we trained on all 8,593 rotated and frontal face images in the training set and tested on the whole test set. This experiment contained four different tests: Global approach using one linear SVM classifier for each person, using one linear SVM classifier for each cluster, using one second degree polynomial SVM classifier for each person, and component-based approach using one linear SVM classifier for each person.

![Figure 6. ROC curves when trained and tested on frontal and rotated faces.](image)

In the second experiment we trained only on the 1,383 frontal face images in the training set but tested on the whole test set. This experiment contained three different tests: Global approach using one linear SVM classifier for each person, using one linear SVM classifier for each cluster, and component-based approach using one linear SVM classifier for each person.

The ROC curves of these two experiments are shown in Fig. 6 and Fig. 7, respectively. Each point on the ROC curve corresponds to a different value of the classification threshold $t$ from Eq. (3). At the end points of the ROC curves

---

5 Before extracting the components we applied the same preprocessing steps to the detected $40 \times 40$ face image as in the global systems.
the rejection rate is 0. Some results of the component-based recognition system are shown in Fig 8.

There are three interesting observations:

- In both experiments the component system clearly outperformed the global systems. This although the face classifier itself (5 linear SVMs) was less powerful than the classifiers used in the global methods (5 non-linear SVMs in the global method without clustering, and 20 linear SVMs in the method with clustering).

- Involving clustering lead to a significant improvement of the global method when the training set included rotated faces. This is because clustering generates viewpoint-specific clusters that have smaller in-class variations than the whole set of images of a person. The global method with clustering and linear SVMs was also superior to the global system without clustering and a non-linear SVM (see Fig. 6). This shows that a combination of weak classifiers trained on properly chosen subsets of the data can outperform a single, more powerful classifier trained on the whole data.

- Adding rotated faces to the training set improves the results of the global method with clustering and the component method. Surprisingly, the results for the global method without clustering got worse. This indicates that the problem of classifying faces of one person over a large range of views is too complex for a linear classifier. Indeed, the performance significantly improved when using non-linear SVMs with second-degree polynomial kernel.

6. Conclusion

We presented a component-based technique and two global techniques for face recognition and evaluated their performance with respect to robustness against pose changes. The component-based system detected and extracted a set of 10 facial components and arranged them in a single feature vector that was classified by linear SVMs. In both global systems we detected the whole face, extracted it from the image and used it as input to the classifiers. The first global system consisted of a single SVM for each person in the database. In the second system we clustered the database of each person and trained a set of view-specific SVM classifiers.

We tested the systems on a database which included faces rotated in depth up to about 40°. In all experiments the component-based system outperformed the global systems even though we used more powerful classifiers (i.e. non-linear instead of linear SVMs) for the global system. This shows that using facial components instead of the whole face pattern as input features significantly simplifies the task of face recognition.

Acknowledgements

The authors would like to thank V. Blanz, T. Serre, and V. Schmid for their help. The research was partially sponsored by the DARPA HID program. Additional support was provided by a grant from Office of

References


