Face Recognition with Shape Features

THESIS

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the
Graduate School of The Ohio State University

By
Xiaojing Xu
Graduate Program in Electrical and Computer Science

The Ohio State University
2015

Master's Examination Committee:
Professor Aleix Martinez, Advisor
Professor Ness Shroff
Abstract

Face recognition has become more significant in recent years. Facial landmarks, on the other hand, have been proven to be useful in face recognition. In this paper, we propose a method to identify frontal view faces using facial landmarks, and show that it is robust to 2D rotation, change of scale and change of position. We propose a face descriptor revealing the shape features, the Euclidean distance between landmarks and their mean. We evaluate the effectiveness of the representation by using it to perform face classification on our emotion dataset. We fit a multivariate normal distribution for each identity and construct a Naïve Bayes Classifier for classification. We show that our method performs well with a small classification error.
Dedication

This document is dedicated to my family, Shaoyuan Xu, Li Luo and Jingrong Liu
Acknowledgments

I would like to express my gratitude to my advisor, Prof. Aleix Martinez, for his continuous inspiration, assistance and guidance. His expertise in computer vision, machine learning, and cognitive science added considerably to my knowledge and experience.

I am also grateful to the other committee member, Prof. Ness Shroff, who has always been kind and supportive, and has become the model of a good professor.

I would like to thank Fabian Benitez-Quiroz, Ruiqi Zhao, Pamela Pallett, Yueyang Zou and all in Computational Biology and Cognitive Science Lab, for their direct and indirect inputs to this thesis. Especial thanks goes to my friends Chang Xu and Qiran Chen who helped me through my hardest time.

Finally, my sincere thanks goes to my parents and my grandmother. Without their support, encouragement and understanding this thesis would not have been possible.
Vita

2014 ................................................................. B.S. Engineering, Shanghai Jiao Tong University

2013 to present ................................................. Department of Electrical and Computer

Engineering, The Ohio State University

Fields of Study

Major Field: Electrical and Computer Engineering
Table of Contents

Abstract............................................................................................................................................ ii

Dedication....................................................................................................................................... iii

Acknowledgments........................................................................................................................... iv

Vita................................................................................................................................................... v

Fields of Study.................................................................................................................................. v

Table of Contents............................................................................................................................ vi

List of Tables .................................................................................................................................. vii

List of Figures ................................................................................................................................ viii

Chapter 1: Introduction .................................................................................................................. 1

Chapter 2: Proposed Method .......................................................................................................... 4

Chapter 3: Experimental Results .................................................................................................... 11

Chapter 4: Conclusion .................................................................................................................... 14

References ..................................................................................................................................... 15
List of Tables

Table 1. Comparison of dimensions and types of covariance matrix............................................ 12

Table 2. Classification errors.......................................................................................................... 13
List of Figures

Figure 1. Dataset.............................................................................................................................. 4

Figure 2. 78 shape features.............................................................................................................. 5

Figure 3. Fitting a Gaussian Model ................................................................................................ 8
Chapter 1: Introduction

Due to its significant applications, automatic face recognition is attracting more attention in recent years. The recognition of faces is an essential ability of human, but it is still hard for automatic face recognition systems to perform as well as human under different conditions, including occlusion, variation of poses, expressions, illumination and etc. [1].

The pipeline of face recognition generally consists of four steps [1]. Face detection finds the subarea in the image that contains the face. Face alignment positions the face detected into a canonical pose, usually represented by a model or target face. Face representation describes the face with certain aspects of interest, i.e. descriptors. And face classification decides whether the representation belongs to a model or target face.

In this paper, the term face recognition refers to only the last two steps, representation and classification. The representation step is very important, because without a good descriptor of the face, little information is fed into the classifier, and thus even the best classifier is not able to give good results. A good representation maximizes the difference between different classes, while keeping the similarity among the same class. Based on face representation, there are mainly two methods of face recognition: appearance-based methods and feature-based methods [2], [3], [4]. Appearance-based methods, such as eigenfaces, extract statistical features. Feature-based methods, such as the EBGM (Elastic Bunch Graph Matching) method
[5], make use of location and statistics of local features such as eyes, nose and mouth.

Facial landmarks capture geometric attributes of human faces, and indicate the positions of informative facial local features that can serve for the purpose of discriminating one class of faces from other faces or non faces. Facial landmarks can be used in face alignment, and both appearance-based and feature-based methods Facial landmark detection has been extensively studied. Typical methods are Active Appearance Model (AAM) – based methods, constrained local model (CLM) – based methods, regression based methods and other methods [6].

In classification step, a representation of the face is compared to the representation of a target face or a model based on specific rules. Frequently used classifiers are Logistic Regression, Bayes Classifier, SVM, Decision Tree and Neural Network.

In this paper, we derived a feature-based method of face representation. It uses landmarks, i.e. the coordinates of fiducial points marking eyes, nose, mouth, eyebrows and the outline of face, to extract descriptor of the face. Our paper doesn’t explore the facial landmark detection process, but focuses on how to use its results. More precisely, we use manually marked facial landmarks to construct the descriptor and perform classifications.

Based on landmarks, we use the Euclidian distances between landmarks and their mean as features to form a representation of the face. The representation proves to be efficient with good classification performance using one of the simplest classifiers, Naïve Bayes Classifier, which requires the distributions of feature vectors from each class. We fit a Gaussian distribution to feature vectors in each class.
Our method is based on the following three assumptions: (1) the faces are viewed from the front, so that our method doesn’t consider varying point of view of faces; (2) the landmarks precisely indicate the position of specific facial features, that is, our method doesn’t deal with variation in resolution of the image, illumination and occlusion; (3) the feature vectors for each identity are Gaussian distributed in the feature space, and vectors from different identities follow different Gaussian distribution. Our choice of representation of faces has the properties of 2D rotation invariance and position invariance by definition, and by normalization we make it scale invariance.
Chapter 2: Proposed Method

In this chapter, we describe our method using our Emotion Database as an example. It contains 5980 frontal view faces of 230 people, each with 26 images of different facial expressions. Faces are manually marked with 78 landmarks. (Figure 1)

Figure 1. Dataset

2.1 Features

For each image of face, we use the distances between landmarks and their mean to form a
feature vector with a length of 78 (Figure 2), their mean meaning the average position of the 78 landmarks, thus forming a 78 dimensional vector space. Each vector in this vector space corresponds to a combination of the relative positions of landmarks. The distances will not change if the face is rotated by different angles or shifted to different positions in the image, so our representation of face is intrinsically rotation and shift invariant. To make it also robust to scale change, we normalize the feature vector before subsequent process.

![Figure 2. 78 shape features](image)

### 2.2 Distribution fitting

A same person’s face can present different features in images because of different facial expressions and poses. For example, when people feel fear, their brows raise, which results in a larger distance from landmarks on the brows to the mean. On the other hand, the feature vectors of the same person will not deviate too far from the mean. Thus we assume that the feature vectors of a same person follow a multivariate normal distribution in the feature space.
The normal distribution \( P(x_i \mid \theta_k) \) of one variable \( x_i \) given one class \( \theta_k \), parameterized by the mean \( \mu_{ki} \) and the variance \( \sigma_{ki}^2 \), is

\[
P(x_i \mid \theta_k) = \frac{1}{\sqrt{2\pi \sigma_{ki}^2}} \exp\left(-\frac{(x_i - \mu_{ki})^2}{2\sigma_{ki}^2}\right).
\] (2.2.1)

And the multivariate normal distribution \( P(x \mid \theta_k) \) of vectors given a class \( \theta_k \), parameterized by the mean vector \( \mu_k \) and the covariance matrix \( \Sigma_k \), is

\[
P(x \mid \theta_k) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} \exp\left(-\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right).
\] (2.2.2)

Where \( d \) is the number of dimensions of the feature vector.

If we assume that the features are mutually independent, then the covariance matrix is diagonal, and (2.2.2) becomes

\[
P(x \mid \theta_k) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi \sigma_{ki}^2}} \exp\left(-\frac{(x_i - \mu_{ki})^2}{2\sigma_{ki}^2}\right).
\] (2.2.3)

In our case, the number of samples, 26, is smaller than the feature dimension, 78, so it is unstable when the covariance matrix is not diagonal, i.e., when we assume that the correlation between two different features is useful in discriminating between two classes.

We proposed two approaches to solve this problem. The first is dimension reduction before fitting the distribution, and assume a full covariance matrix. PCA serves well for this purpose.
The second approach assumes that the covariance matrices are diagonal. As a result, the number of samples doesn’t matter much, unless the situation that some of them lies on the same line. However, both approaches have the problem of possible overfitting. These parameters are optimized by choosing the dimension as well as the type of the covariance matrix with smallest leave-one-expression-out error, which is introduced later.

It should be noted that when fitting the distribution, it turns out that for each person the Gaussian distribution model with one component has the smallest AIC, meaning that the distribution of feature vectors of one person is clustered in the vector space, and can be well-modeled by a one-component multivariate Gaussian distribution. This can also be proved by comparing the leave-one-expression-out errors with different number of Gaussian components.
After dimension reduction

Gaussian contours

Figure 3. Fitting a Gaussian Model

2.3 Bayes Classification

With distributions of different classes known, we design a Bayes Classifier. Bayes Classifier is based on the following Bayes decision rule:
\[
\hat{\theta}(x) = \arg\max_{y \in \{\theta_1, \theta_2, \ldots, \theta_n\}} P(y \mid x).
\] (2.3.1)

Where \(x\) is the observation, \(\theta_1, \theta_2, \ldots, \theta_n\) are the \(n\) classes, in our case \(n=230\), and usually we set \(\theta_n = n\). \(\hat{\theta}(x)\) is the estimated class based on the observation \(x\). \(P(y \mid x)\) is called the posterior probability. Bayes rule decides on the class that has the largest posterior probability. Based on the Bayes' rule, the posterior probability

\[
P(y \mid x) = \frac{P(y)P(x \mid y)}{P(x)}. \tag{2.3.2}
\]

And thus (2.3.1) becomes,

\[
\hat{\theta}(x) = \arg\max_{y \in \{\theta_1, \theta_2, \ldots, \theta_n\}} \frac{P(y)P(x \mid y)}{P(x)}. \tag{2.3.3}
\]

\(P(y) \ (y \in \{\theta_1, \theta_2, \ldots, \theta_n\})\) is called the prior probability. In our case, the prior probabilities for different classes are equal, and the denominator, \(P(x)\) is constant for each class, so in the end,

\[
\hat{\theta}(x) = \arg\max_{y \in \{\theta_1, \theta_2, \ldots, \theta_n\}} P(x \mid y). \tag{2.3.4}
\]

That is, when the prior probabilities are the same, Bayes classifier assign the observation to the class for which the conditional probability of the observation conditioning on the class is the greatest.

In our case, given an observation \(x\), we know \(P(x \mid y)\) for each \(y\) directly from the distribution
of y. Thus it is easy to determine which class it belongs to.

When the covariance matrix for the distribution of each identity is diagonal, the Bayes Classifier is actually a Naïve Bayes Classifier. Naïve Bayes Classifiers assume that the features are strongly independent, and thus, the probability of the intersection of the events is the product of the probabilities of all the events. When the data follow a jointly Gaussian distribution, uncorrelation indicates strong independence, so the model becomes a Naïve Bayes Model, and (2. 3.4) becomes

$$\hat{\Theta}(x) = \arg\max_{\theta \in \{\theta_1, \theta_2, \ldots, \theta_n\}} \prod_{i=1}^{d} P(x_i \mid y)$$  \hspace{1cm} (2.3.5)

Where d is the dimension of the feature vector, and \( x_i \) is the i\(^{th} \) feature in the feature vector \( x \).

The Naïve Bayes classifier assigns the observation to the class for which the product of the probabilities of features is the greatest.
3.1 Evaluation

Two ways are used to evaluate the method: (1) Training error: classify all the samples in the training set, calculate the misclassification rate; (2) Leave-one-expression-out error: leave one facial expression of all identities out, train the classifier using the remaining samples, and test on the left out samples. Repeat for each facial expression, and calculate the mean error rate.

3.2 Choice of dimension and covariance matrix

As mentioned above, there are two things we should consider when fitting the distribution. One is whether to assume the features to be independent, another is how many dimensions to choose to fit the distribution. We use the leave-one-expression-out error to evaluate the method, and choose the combination which results in the least error. We use grid search to find the ideal combination. The results are shown in table 1 (only points near the optimum are shown).
<table>
<thead>
<tr>
<th>Covariance matrix</th>
<th>diagonal</th>
<th>full</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of dimensions</td>
<td>67 66 65 25 24 23</td>
<td></td>
</tr>
<tr>
<td>Error rate</td>
<td>9.53% 9.38% 9.40% 13.26% 13.85% 15.50%</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Comparison of dimensions and types of covariance matrix

It can be seen from the table that with more features, diagonal covariance matrix preserve more information and thus results in a lower error rate. And with diagonal covariance matrix, 66 dimensions lead to the lowest error rate. This means that for our method on our dataset, 66 is the best number of dimensions to choose to preserve covariance while preventing overfitting.

### 3.3 Results

Except for the number of dimensions and type of covariance matrix discussed above, we also compare the leave-one-expression-out errors with different number of components in the Gaussian Model, and the result show that one component is the best.

At last, the training error and leave-one-expression-out error are as follows:
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Training error</td>
<td>1.17%</td>
</tr>
<tr>
<td>Leave-one-expression-out error</td>
<td>9.38%</td>
</tr>
</tbody>
</table>

Table 2. Classification errors
Chapter 4: Conclusion

Shape features are significant in face recognition. Our work suggests that the landmarks can be directly used in face classification. The features we use, the Euclidean distance between landmarks and their mean, have the properties of rotation and shift invariance, and since our method involves normalization, it is also robust to scale change. We believe that combined with other features, or more complex classifiers, the shape features introduced by landmarks will have very useful application in face recognition.
References


