

# 基于 PCA 的特征选择算法

于成龙

(南京邮电大学 计算机学院, 江苏 南京 210003)

**摘要:**在人脸识别的某些应用中,最好能够找到原始特征的关键子集,减少不必要的特征计算和资源耗费,而不是得到所有原始特征的映射。主成分分析法(Principal Components Analysis, PCA)是目前比较常用的人脸识别算法,PCA 将人脸图像映射到能很好地表征训练图像集的特征脸空间中,但是基于 PCA 的人脸识别的缺陷在于原始空间所有的特征都映射到了低维特征空间中,是基于最佳描述性特征子集。提出了一种新的基于 PCA 的特征选择方法,将特征选择与特征抽取相结合,对特征脸空间再进行特征选择,选择人脸原始特征集中最关键的特征,并将其应用在基于 PCA 的人脸识别中。

**关键词:**人脸识别; PCA; 特征脸; 特征选择

**中图分类号:**TP301.6

**文献标识码:**A

**文章编号:**1673-629X(2011)04-0123-03

## Features Selection Algorithm Based on PCA

YU Cheng-long

(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

**Abstract:** In some applications of face recognition, it might be more desirable to pick a subset of the original features than to find a mapping that uses all of the original features. The benefits of finding this subset of features lie in cost reduced computations and thus lower cost of sensors. Principal components analysis (PCA) is widely used in face feature extraction and recognition. The facial images are projected onto eigenfaces that best define the variation of the known test images. However, the PCA-based face recognition has the disadvantage that, on the basis of an optimal descriptive feature subset, measurements from all the original features are used in the projection to the lower dimensional space. Propose a new method for dimensionality reduction of a feature set by choosing a subset of original features that contains most of the essential information. This method, based on PCA, combines together feature selection and feature extraction. The proposed method has been successfully applied in choosing principal features in PCA-based face detection and recognition.

**Key words:** face recognition; PCA; eigenface; feature selection

## 0 引言

特征提取和特征选择是人脸识别中数据预处理阶段的关键技术。特征提取是将原始特征进行某种形式的变换以得到新的特征。特征选择依据某种评估准则,从原始特征集中选择最优的特征子集。

在人脸识别的研究中,尽管通过 PCA 算法可获得特征脸空间,但特征脸是原始人脸中所有原始特征的映射,相对于某些应用要求,并非符合最优标准下保持原始数据中大部分的相关信息的要求。文中用一种基于 PCA 的特征选择方法<sup>[1,2]</sup>,试图先得到特征脸空间,然后使用 KNN 聚类的方法<sup>[3,4]</sup>对特征脸保留的原始特征再进行特征选择,得到进一步约减的特征脸,以达

到用特征选择和特征抽取相结合的方法来进行人脸识别的目的。

## 1 PCA 算法

假设一个给定的训练数据集含有  $N$  个样本  $X \in R^n$ , 每个样本由  $n$  维特征向量描述为:

$$X_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T \in R^n, \text{均值为 } m。$$

训练集的协方差矩阵定义为:

$$\Sigma = \sum_{i=1}^n (X_i - m)(X_i - m)^T \quad (1)$$

$\Sigma$  的特征值表示样本在特征矢量上的分布方差。

$$A = \begin{pmatrix} \hat{e}_1 & & 0 & \hat{u}_1 \\ \hat{e}_2 & & & \hat{u}_2 \\ \hat{e}_3 & & 0 & \hat{u}_3 \\ \hat{e}_n & & & \hat{u}_n \end{pmatrix}, \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$$

选择  $\Sigma$  的  $d$  个特征矢量根据特征值进行排序,降维后的特征子空间表示为  $Y \in R^d, d < n$ :

$$Y_i = Q^T X \quad (2)$$

收稿日期:2010-07-31;修回日期:2010-11-04

基金项目:江苏省自然科学基金(08KJB520008);南京邮电大学人才引进启动基金(NY207137, NY207148)

作者简介:于成龙(1984-),男,硕士研究生,研究方向为模式识别与图像处理。

其中  $Y_i = \{y_{i1}, y_{i2}, \dots, y_{id}\}^T$ 。

PCA 最重要的特性是使得样本在低维空间中尽量分散、保留样本在原始空间中的差异性和在低维空间中的投影数据与原始数据之间的均方差最小<sup>[3,5]</sup>。

## 2 Eigenface 算法

特征脸方法将包含人脸的图像区域看作是一个随机向量,经过 PCA 变换,对应其中较大特征值的基底具有与人脸相似的形状,因此又称特征脸<sup>[6]</sup>。利用特征脸的线性组合可以描述、表达和逼近原始人脸图像,从而进行人脸识别与合成。识别过程就是将待识别的人脸图像映射到由特征脸构成的子空间上,比较其与已知人脸在此特征子空间中的位置,具体步骤如下:

(1) 初始化获得人脸图像集,进行 PCA 变换,选择保留的特征向量个数:

$$\text{Retained} = \frac{\sum_{i=1}^d \lambda_i}{\sum_{i=1}^n \lambda_i} \cdot 100\% \quad (3)$$

保留的一个特征向量,就是一个特征脸,保留的所有特征向量,构成特征脸空间;

(2) 输入训练人脸,将其映射到特征脸空间中,得到一组关于训练集的特征数据;

(3) 输入测试人脸,将其映射到特征脸空间中,得到关于该测试人脸的特征数据;

(4) 计算测试人脸与训练人脸之间的距离或相似性,进行识别。

## 3 基于 PCA 的特征选择

假设人脸  $X_i$  在第  $j$  个主成分的投影为:

$$Y_i = Q^T X \quad (4)$$

$$\text{即 } y_{ji} = q_j^T x_i = \sum_{k=1}^n q_{jk} x_{ik} \quad (5)$$

如公式(5),人脸图片在特征脸空间上的投影是所有原始特征的线性组合。但有些特征可能是冗余的或没有意义的。通过特征选择可以找到那些在应用中起重要作用的关键特征,摒弃冗余的特征<sup>[7,8]</sup>。

特征  $x_{ik}$  的意义可以通过变换矩阵中与其相应的参数  $q_{jk}$  来评价,即根据变换矩阵中的元素来确定与特征脸关系密切的关键原始特征<sup>[9]</sup>。另一方面,变换矩阵可以表示为:

$$Q = \{V_1, V_2, \dots, V_d\}, V_i = \{q_{i1}, q_{i2}, \dots, q_{id}\}$$

$$i = 1, 2, \dots, n$$

向量  $|V_1|, |V_2|, \dots, |V_n| \in \mathbb{R}^d$  被称为行成分,表示第  $i$  个原始特征在低维特征脸空间上的投影,即行成分中的  $d$  个观察值是特征脸子空间中的各个坐标

(主成分)上的投影权重<sup>[3]</sup>,原始特征与行成分是一一对应的。如果原始特征关联性很强,其在子空间上的投影权重也会非常相近,即具有类似的行成分。在极端情况下,对于两个相互独立的特征,它们的投影权重有极大的不同;而两个完全关联的特征,它们有着相同的投影权重(不考虑符号因素)<sup>[10]</sup>。基于以上的观察,可以通过选择行成分,而与所选择的行成分对应的原始特征就是最终所选择的最优特征子集。

为了寻找特征子集,特征选择方法是利用行成分的结构特性,通过聚类使得子集中的行成分高度关联<sup>[2,11]</sup>,然后从每一个子集中选出一个代表性的行成分。所选出的行成分能很好地代表其所在子集中的所有行成分。而与代表性行成分对应的原始特征就是选择出来的特征,所选出的特征数量与聚类(子集)的数目一致,从而实现对特征脸进行特征选择,得到新的特征脸空间。特征选择算法的简单流程如下:

a) 利用 PCA 获得变换矩阵  $Q$ ,生成特征脸空间,保留的特征向量数目为:

$$\text{Retained} = \frac{\sum_{i=1}^d \lambda_i}{\sum_{i=1}^n \lambda_i} \cdot 100\% \quad (6)$$

b) 提取  $Q$  矩阵行成分  $|V_1|, |V_2|, \dots, |V_n| \in \mathbb{R}^d$  两两比较,构建相似性矩阵,距离度量为欧式距离;

c) 使用 KNN 聚类方法将行成分  $|V_1|, |V_2|, \dots, |V_n| \in \mathbb{R}^d$  聚成  $p$  个类<sup>[3]</sup>;

d) 从每个聚类中找到与类中心最近的行成分  $V_i$ ,相应的原始特征就是关键特征,最终将选出  $p$  个关键特征,生成约减后的特征脸空间<sup>[4,12]</sup>。

## 4 行成分相似性度量

如何计算行成分之间的相似性,也是文中考虑的一点。常用的计算变量之间相似性的方法有关联系数、欧式距离以及最大信息压缩(MICI)<sup>[8]</sup>等。Mittra 的分析和实验已表明 MICI 优于其它方法<sup>[7]</sup>。现简单介绍如下:

对于行成分  $|V_1|, |V_2|, \dots, |V_n| \in \mathbb{R}^d$  其两两相似性度量如下:

$$\text{MICI: } 2\lambda(q_i, q_i) = D(q_i) + D(q_i) - \sqrt{(D(q_i) + D(q_i))^2 - 4D(q_i)D(q_i)(1 - \rho_{qq_i}^2)} \quad (7)$$

$$\text{欧式距离: } D(q_i, q_i) = \sqrt{(q_i - q_i)^2} \quad (8)$$

其中  $D$  表示变量的方差,  $\rho$  表示关联系数。

## 5 实验结果

在缺少先验知识的情况下,无法预先知数据集中各个特征的重要性,为了验证算法的性能,文中通过

与传统PCA方法在分类性能以及原始人脸重建效果上进行对比实验。

整个实验包括两个部分:首先,选择基准ORL人脸库,输出PCA特征脸以及特征选择后的特征脸,然后分别用于原始人脸重建;另外在人脸识别方面比较基于PCA的特征选择算法(以下简称为PFA)<sup>[5]</sup>和PCA的性能。

实验一如图1所示,PCA特征脸感官上和整体上近似于原始人脸,重建的原始人脸具有最佳描述性的特点;而PFA特征脸,摒弃了原始人脸的冗余特征,再次降维后将特征更集中于眼睛、鼻子、嘴巴以及人脸的主要轮廓,其重建后的人脸灰度值比较集中,集中重建了原始人脸的轮廓和五官<sup>[5]</sup>。



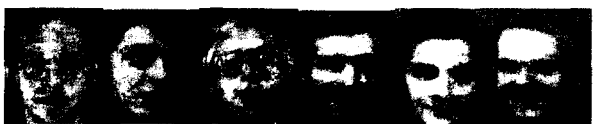
(a) 原始人脸



(b) PCA特征脸



(c) PFA特征脸



(d) PCA特征脸重建效果



(e) PFA特征脸重建效果

图1 PCA特征脸与PFA特征脸实验比较

实验二是在ORL标准数据库上,使用KNN聚类对特征脸进行特征选择,并比较PCA、PFA+欧式距离和PFA+MICI的识别率。结果如图2所示。

约减的特征脸,有降低时间和运算开销,避免维数灾难等意义。由实验得知,特征选择后的特征脸,更集中突出了人脸的眼睛、鼻子、嘴巴等五官特征以及人脸整体轮廓等主要特征<sup>[12]</sup>,从而实现了将PCA特征脸与特征选择相结合来进行人脸识别的预处理的目的。

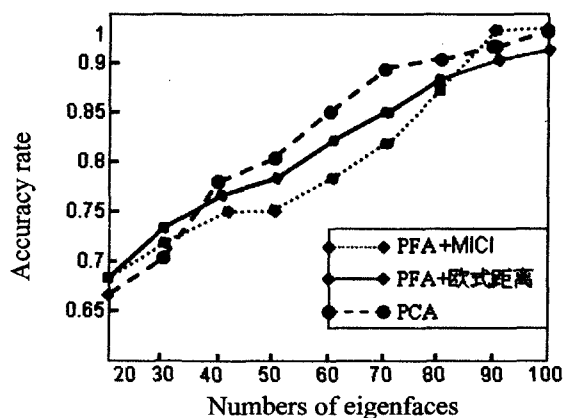


图2 人脸识别率函数图

#### 参考文献:

- [1] Jolliffe I T. Principal Component Analysis [M]. New York: Springer-Verlag, 1996.
- [2] 陈彬,洪家荣,王亚东. 最优特征子集选择问题[J]. 计算机学报, 1997, 20(2): 17-22.
- [3] McCabe G P. Principal Variables [J]. Technometrics, 2004, 26: 134-137.
- [4] 张莉,孙钢,郭军. 基于K-均值聚类的无监督的特征选择方法[J]. 计算机应用研究, 2005(3): 23-24.
- [5] 张洁,高新波,焦李成. 基于特征加权的模糊聚类新算法[J]. 电子学报, 2006, 34(1): 412-420.
- [6] Turk M A, Pentland A P. Face recognition using eigenfaces [C] // Proc. CVPR. [s. l.]: IEEE, 2001: 586-591.
- [7] 李云,叶春晓. 基于特征关联性的特征选择算法研究[J]. 微型机与应用, 2004(6): 58-60.
- [8] 王嘉驹. 复杂数据的特征选择与关联分析[D]. 上海: 上海交通大学, 2005.
- [9] 范劲松,方廷建. 特征选择和提取要素的分析及其评价[J]. 计算机工程与应用, 2001, 37(13): 95-99.
- [10] Cao L, Miao Y M. Exction interactions in CdS nanocrystal aggregates in reverse micelle [J]. J. Chem. Phys., 2005, 123: 24-30.
- [11] Molina L C, Belanche L, Nebot A. Feature selection algorithm: a survey and experimental evaluation [C] // In: Proc. 2002 IEEE International Conference on Data Mining. [s. l.]: [s. n.], 2002: 306-313.
- [12] Dash M, Liu H. Feature selection for clustering [C] // Proc. of Fourth Pacific Asia Conf. on Knowledge Discovery and Data Mining. [s. l.]: [s. n.], 2000: 110-121.

## 6 结束语

基于PCA的特征选择算法是一种有效的、具有实际意义的人脸图像处理和人脸识别方法。文中在基于PCA的基础上,再对PCA特征脸进行特征选择,摒弃特征脸中冗余的特征,选择更有效的特征,得到进一步