# 基于样本选择的最近邻凸包分类器

姜文瀚 周晓飞 杨静宇
(南京理工大学计算机科学与技术学院, 南京 210094)

**摘 要** 最近邻凸包分类算法是一种以测试点到各类别样本凸包的距离为分类度量的最近邻分类算法。然而, 该算法的凸二次规划问题优化求解的较高的计算复杂度限制了其在较大规模数据集上的应用。本文提出一种样本选择方法——子类凸包生长法。通过迭代, 选择距离选出样本凸包最远的点, 直到满足终止条件, 从而实现数据集的有效约简。ORL数据库和 MIT-CBCL人脸识别 training-synthetic库上的实验结果表明, 子类凸包生长法选出的少量样本生成的凸包能够很好的表征训练集, 在不降低最近邻凸包分类器性能的同时, 使得算法的计算速度大为提高。

**关键词** 样本选择 凸包 最近邻凸包分类 子类凸包生长
**中图法分类号**: TP391 **文献标识码**: A **文章编号**: 1006-8961(2008)01-0109-05

# A Nearest Neighbor Convex Hull Classifier with Sample Selection

JIANG Wen-han, ZHOU Xiao-fei, YANG Jing-yu
*(Department of Computer Science and Technology, Nanjing University of Science and Technology of China, Nanjing 210094)*

**Abstract** The nearest neighbor convex hull(NNCH) classification algorithm is a kind of nearest neighbor classification method which takes the approximation errors of the convex hulls of all members of every class to the test point as the discriminant measures. However, the higher computation costs of quadratic optimization problems of the algorithm limit its applications on large data sets. So a sample selection method for NNCH named subclass convex hull growth is proposed in this paper. For one class data, the farthest two points are selected first as the initial chosen set. Then, the distances of others to the convex hull of the chosen set are computed respectively. We choose the farthest one and add it into the chosen set. This procedure is repeated until the end conditions. The convex hull of selected samples is taken as the approximation of all. The more samples are selected, the less approximation error is achieved, so the valid estimation of sample distribution is realized. Experiments on the ORL database and the MIT-CBCL face recognition training-synthetic database show the abilities of this method to reduce the training data and accelerate the computation while maintaining the generalization performance of NNCH.

**Keywords** samples selection, convex hull, nearest neighbor convex hull(NNCH), classification, subclass convex hull growth

## 1 引 言

最近邻凸包(nearest neighbor convex hull, NNCH)分类器是一种以测试点到各类别的所有训练样本生成凸包距离为相似性度量的最近邻分类算法。文献[1]中, NNCH分类器表现了与支持向量机相当的分类性能。然而, 最近邻凸包分类算法的点到凸包距离计算需要求解一个凸二次规划(quadratic programming, QP)问题。QP问题的标准求解优化算法(如 Matlab OP例程)的时间复杂度是 $O(M^3)$[2] (NNCH中, $M$ 是一类训练样本的个数)。

**基金项目**: 国家自然基金资助项目(60472060)
**收稿日期**: 2006-06-27; **改回日期**: 2006-09-25
**第一作者简介**: 姜文瀚(1974~ ), 男。现为南京理工大学计算机科学与技术学院博士研究生。主要研究方向为人工智能与模式识别。
E-mail: wenhan@njust.edu.cn

QP方程的 Gram矩阵大小为 $M \times M$。显然, 与训练集规模密切相关的计算复杂度限制了最近邻凸包分类算法的大数据集应用。一个直接有效的解决办法就是在使分类性能损失较小的前提下, 以适当的选择策略减少训练集样本。

因此, 本文为 NNCH 分类器设计了一种样本选择方法。在一类训练样本集中, 用选择的较少的凸包边缘点生成的子类凸包 (由一类训练样本的子集生成的凸包称为子类凸包) 来估计该类别样本分布。该方法是一种只与同类样本分布有关, 而与异类无关的样本选择方案。

## 2 最近邻凸包分类

最近邻凸包分类算法是一种利用凸包估计样本分布, 并以凸包为原型的最近邻分类方法。该方法假定同类样本会分布在同一凸包里或离该凸包较近的区域, 并将凸包作为模式分布的一种粗略估计。从凸包分布估计的角度分析, 支持向量机的直观几何意义[3]也是最大间隔地实现训练样本凸包的相互分离。

作为样本相似性度量依据的点到凸包距离的计算是分类算法的核心。对于含有 $k$ 个样本的第 $i$ 类训练样本集 $\boldsymbol{S}_i = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_k\}$, $\boldsymbol{S}_i \subset \mathbf{R}^n$, $\boldsymbol{S}_i$ 的凸包为

$$co(\boldsymbol{S}_i) = \left\{ \boldsymbol{x} = \sum_{j=1}^{k} \lambda_j \boldsymbol{x}_j \;\middle|\; \sum_{j=1}^{k} \lambda_j = 1,\ \lambda_j \geqslant 0,\ j = 1, \cdots, k \right\}$$

在此定义由一类全部训练样本生成的凸包为类凸包。

样本点 $\boldsymbol{x} \in \mathbf{R}^n$ 到 $\boldsymbol{S}_i$ 的类凸包 $co(\boldsymbol{S}_i)$ 的距离为

$$d(\boldsymbol{x}, co(\boldsymbol{S}_i)) = \inf_{\boldsymbol{y} \in co(\boldsymbol{S}_i)} \|\boldsymbol{x} - \boldsymbol{y}\| \tag{1}$$

令 $\alpha = (\alpha_1, \cdots, \alpha_k)^{\mathrm{T}}$, $\boldsymbol{y}$ 是类凸包 $co(\boldsymbol{S}_i)$ 上的点, 根据凸包定义, 式 (1) 可化为

$$d(\boldsymbol{x}, co(\boldsymbol{S}_i)) = \min_{\alpha} \left\| \boldsymbol{x} - \sum_{i=1}^{k} \alpha_i \boldsymbol{x}_i \right\|$$
$$\text{s.t.} \quad \sum_{i=1}^{k} \alpha_i = 1,\ \alpha_i \geqslant 0 \tag{2}$$

在欧式距离定义下有

$$d^2(\boldsymbol{x}, co(\boldsymbol{S}_i)) = \min_{\alpha} \left\| \boldsymbol{x} - \sum_{i=1}^{k} \alpha_i \boldsymbol{x}_i \right\|_2^2$$
$$= \min_{\alpha} [\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} - 2\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k)\alpha + \alpha^{\mathrm{T}}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k)^{\mathrm{T}}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k)\alpha] \tag{3}$$
$$\text{s.t.} \quad \sum_{i=1}^{k} \alpha_i = 1,\ \alpha_i \geqslant 0$$

式 (3) 省略优化无关项 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$, 并除以 2, 得到凸二次规划优化式:

$$\min_{\alpha} \frac{1}{2}\alpha^{\mathrm{T}}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k)^{\mathrm{T}}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k)\alpha - \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k)\alpha$$
$$\text{s.t.} \quad \sum_{i=1}^{k} \alpha_i = 1,\ \alpha_i \geqslant 0 \tag{4}$$

若 $\hat{\alpha} = (\alpha_1, \cdots, \alpha_k)^{\mathrm{T}}$ 是式 (4) 的最优解, 则

$$d^2(\boldsymbol{x}, co(\boldsymbol{S}_i)) = \left\| \boldsymbol{x} - \sum_{i=1}^{k} \alpha_i \boldsymbol{x}_i \right\|_2^2 \tag{5}$$

对于多类问题, NNCH 把测试样本点到各类凸包的距离作为判别依据, 按照最近邻分类原则决定类别归属。即对于 $c$ 类 $\omega = \{\omega_1, \omega_2, \cdots, \omega_c\}$ 问题, 若 $d(\boldsymbol{x}, co(\boldsymbol{S}_j)) = \min_{i=1,\cdots,c} d(\boldsymbol{x}, co(\boldsymbol{S}_i))$, 则 $\boldsymbol{x} \in \omega_j$。

## 3 样本选择

最近邻凸包分类算法涉及的测试点到各类所有样本生成凸包的距离计算需要求解式 (4) 的 *QP* 问题。*QP* 问题标准求解算法的较高的计算复杂度限制了该分类算法的大数据集应用。为此, 本文提出了一种训练样本选择方法。

Krein-Milman定理[4]指出, 一个闭的, 有界凸集是它的极点的凸包。所以, 一类训练样本凸包由其顶点样本就可以生成。从样本选择的角度出发, 只取顶点自然就能够实现样本集约简的目的。然而, 凸包顶点的求解是一个典型的 NP (nondeterministic polynomial) 难题。准确求出所有顶点所耗费的大量计算也是人们所不希望的。既然最近邻凸包分类算法是以凸包作为样本分布估计的, 那么就大可不必严格的求出所有顶点, 只要选出的样本点所生成的凸包能够很好的估计样本分布就可以了。这也是本文样本选择算法的出发点。

针对一类训练样本, 本文首先以该类训练样本集中相距最远的两个样本作为子类凸包 (选出集凸包) 的生成元。在随后的过程中, 不断归并距离该子类凸包最远的点, 直到满足预先设定的退出条件, 见如下算法。

设第 $i$ 类样本集 $\boldsymbol{S}_i$ 包含 $k$ 个样本, 已选样本子集 $\boldsymbol{S}_i'$, $l$ 是已选样本数。

(1) 初始化　设定拟选择样本个数 $m$, 逼近误差界 $\varepsilon$, 初始最大逼近误差 $e_{\max} = \inf$, 初始选择集

$$S_i' = \{x_a, x_b \mid [x_a, x_b] = \arg\max_{x_p, x_q \in S_i} \|x_p - x_q\|_2^2\}$$

(2) 如果选择集 $S_i'$ 的样本个数 $l < m$, 则对于 $\forall x \in S_i \backslash S_i'$, $d(x) = d^2(x, co(S_i'))$, $e_{max} = \max_{x \in S_i \backslash S_i'} d(x)$, $\hat{x} = \arg\max_{x \in S_i \backslash S_i'} d(x)$ ($\hat{x}$ 为距离凸包最远的样本向量); 否则退出。

(3) 如果 $e_{max} > \varepsilon$ 则 $S_i' = S_i' \cup \hat{x}$, 否则退出。

(4) 转至 (2)。

本文形象地称这一过程为子类凸包生长 (subclass convex hull growth, SCHG)法。该过程是通过不断减小的 (子类凸包对样本的 )最大逼近误差 $e_{max}$ 来实现对样本分布的有效估计的。

该算法初始选择的两个生成元是类凸包的顶点; 此后选择归并的距离子类凸包最远 (逼近误差最大 )的点可能不是类凸包顶点, 但一定是边缘点。当最大逼近误差 $e_{max} = 0$时, 实现了子类凸包对类凸包的最佳逼近。对于迭代过程中出现的逼近误差为零的样本, 由于该点在此后的迭代中与子类凸包的距离必为零 (该候选点已在凸包上 ), 因此可直接省去, 以减少计算量。

上述算法可以根据不同的样本选择要求, 设置参数 $m$ 和 $\varepsilon$ 具体分为以下 3种情况:

(1) 以选样个数 $m$ ($2 \leqslant m \leqslant k$)为标准时: 设置 $\varepsilon = -1$, 当选样个数 $l$达到限定个数 $m$ 时选样结束;

(2) 以逼近误差界 $\varepsilon$( $\varepsilon \geqslant 0$)为标准时: 设置 $m = \inf$, 当最大逼近误差 $e_{max} \leqslant \varepsilon$时, 选样结束;

(3) 同时以样本个数和距离界为标准时: 设置相应的 $m$ 和 $\varepsilon$。当选样个数 $l$达到限定个数 $m$ 或最大逼近误差 $e_{max} \leqslant \varepsilon$时, 选样结束。此时各类选择的样本数及逼近误差可能不同。

## 4　人脸识别应用实验及分析

实验首先采用本文的子类凸包生长法进行样本选择, 然后将选出的样本作为新的训练集用于最近邻凸包分类器的分类测试。样本选择以选样个数为标准。算法中涉及的优化计算均由 Matlab7 的 quadprog例程来完成。实验在 Pentum Ⅳ 2.8G, 内存 256M 的 PC机上执行。

首先在剑桥 AT&T 实验室的 ORL( olivetti research lab)标准人脸图像库[5]上对本文方法进行验证。ORL数据库包括 40人, 每人 10幅, 共计 400幅 PGM 格式, 灰度级 256 分辨率 92×112的人脸灰度图像。类别标识为 $s \times$( ×是 1~ 40自然数); 样本标识为 ×( ×是 1~ 10自然数)。图像在不同条件下采集, 如不同时间, 变化光照, 不同表情 (如睁眼 、闭眼 、微笑 、不笑 ), 取舍饰物 (如戴眼镜 、不戴眼镜 ) 等。所有图像均为单一黑色背景, 以前正向脸为主, 兼有些许缩放 、旋转和侧移, 如图 1所示。

图 1　ORL人脸库图像示例

Fig 1　Images from ORL face database

实验中, 全部图像转换为 JPG 格式, 双 3次插值缩至 16×16大小。鉴于该数据库规模较小, 可在全库上进行选样和测试。实验结果如图 2所示。

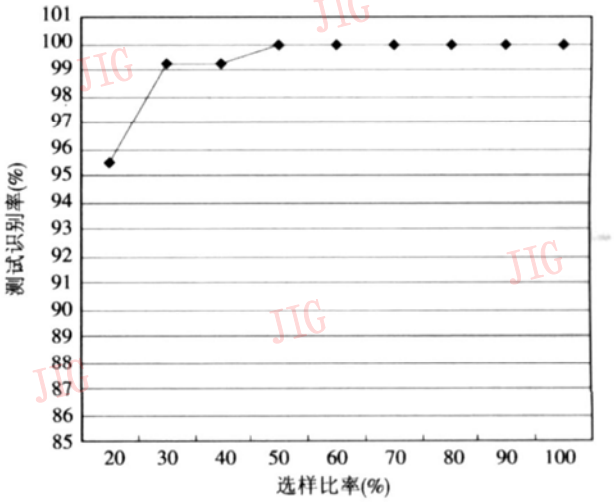



图 2　ORL人脸识别数据库上的选样比率与识别比率

Fig 2　Sample selection and recognition rate on ORL face database

图 2所示图表的横坐标选样比率是选择样本个数与全库样本数之比。可以看出, ORL数据库上的实验, 当每类选择 3个样本时, 测试识别率就已达到 99.25%, 当每类选择 5个训练样本时, 实现了全库样本 100%的正确识别。实验结果说明了本文方法选择样本生成的凸包具有较强的类别表征能力。

另外一组实验是在较大规模的 MIT-CBCL人脸识别数据库[6] (The MIT-CBCL face recognition data-

base)的 training-synthetic库上实现的。该库包含由 3维形态模型合成的姿态和光照变化的 10个人的3 240幅标准人脸图像, 每人 324幅。所有图像仅包含无遮掩椭圆形面颈部区域。姿态变化: 水平左向旋转 0°~32°, 以 4°为增量。光照变化: 以头部为中心, 水平右向 15°~90°, 以 15°为增量; 竖直仰角 0°~75°, 以 15°为增量。图像格式为 PGM, 分辨率为 200×200。如图 3为部分示例图像 (姿态变化 0°, 16°和 32°; 光照变化水平和竖直方向均分别为 15°, 45°和 75°)。

图 3 MIT-CBCL人脸识别 training-synthetic库示例

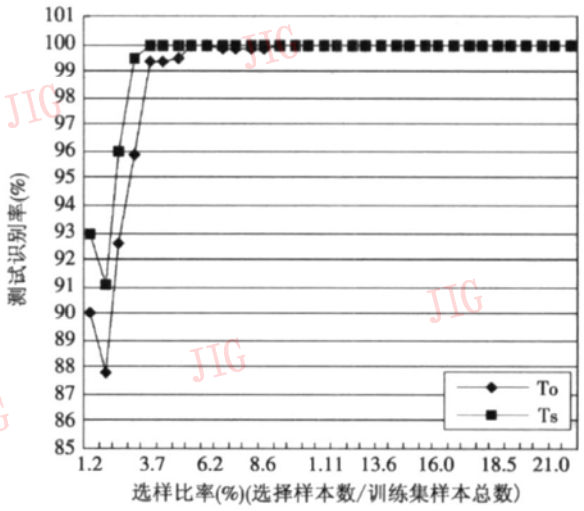
Fig 3 Examples of the MIT-CBCL face recognition training-synthetic database

实验把 training-synthetic库中的人脸图像按照光照仰角分为两个子集 $A_1$ 和 $A_2$, 分别包括仰角为 {0°, 30°, 60°}和 {15°, 45°, 75°}的图像。所以两个子集各包括 10个人的1 620幅图像, 每人 162幅。实验中, 将全部图像统一转换为 JPG格式, 双三次插值缩为 16×16大小。

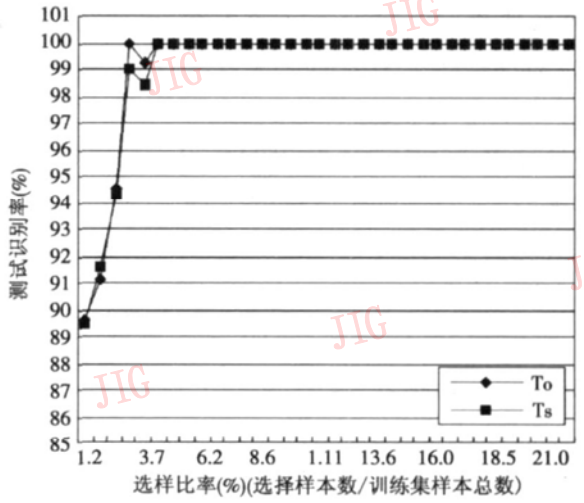
实验首先将 $A_1$ 作为训练集, 从 $A_1$ 中选择样本训练, 并分两种情况进行测试: 一种情况是在训练集 $A_1$ 自身上进行测试, 目的是体现选出样本凸包对训练集类别的表达能力, 测试识别率随选样比率变化的情况如图 4(a)所示的曲线“Ts”; 另一种情况是对 $A_2$ 进行测试, 目的是反映样本选择方法对分类器泛化能力的影响, 见图 4(a)所示的曲线“To”。

同样的实验目的和方法, 实验再将 $A_2$ 作为训练集, $A_2$ 自身测试的结果如图 4(b)所示的“Ts”曲线, $A_1$ 上的测试结果如图 4(b)所示的“To”曲线。

分析图 4(a)的实验数据, 可以看出, 曲线“Ts”当选样比率为 5.6% (各类选择 9个样本)时, 识别率 100%, 选择样本凸包能够完全正确表征训练集类别; 曲线“To”当选样比达到 9.3% (各类选择 15个样本)以后, 测试集识别率也稳定在 100%。分析图 4(b)的实验结果, 曲线“Ts”和曲线“To”在选样比率为 4.9% (每类选择 8个样本)时, 同时达到 100%识别稳定状态。




(a)




(b)

图 4 MIT-CBCL人脸识别 training-synthetic库上的选样比率与识别比率

Fig 4 Sample selection rate and recognition rate on MIT-CBCL face recognition training-synthetic database

实验中, 本文样本选择方法以较少的选择样本即可使得最近邻凸包分类器达到较高的识别比率, 一方面证实了本文的子类凸包生长法所选样本生成的凸包能够有效地表征训练集类别, 另一方面也充分说明了选出的样本能够有效地支持最近邻凸包分类。

另外, 实验分别对以 $A_1$ 作为训练集, 每类选择 15个样本, 测试 $A_2$, 和以 $A_2$ 作为训练集, 每类选择 8个样本, 测试 $A_1$ 的两种情况的执行时间 (包括样本选择时间和测试时间两部分)进行了记录, 并同未经选样直接测试的情况加以比较, 结果如表 1所示, NNCH 表示未经样本选择的最近邻凸包分类算法, SCHG + NNCH 表示进行样本选择的最近邻凸包分类算法。

表 1　样本选择前后最近邻凸包分类器的比较

Tab. 1　NNCH vs NNCH with sample selection

| 训练集 | 实验方法 | 选样数(/类) | 识别率(%) | 选样时间(s) | 测试时间(s) | 合计时间(s) |
|---|---|---|---|---|---|---|
| $A_1$ | NNCH | 162 | 100 | – | 44 660 | 44 660 |
| | SCHG + NNCH | 15 | 100 | 99 | 166 | 265 |
| $A_2$ | NNCH | 162 | 100 | – | 46 125 | 46 125 |
| | SCHG + NNCH | 8 | 100 | 31 | 69 | 100 |

表内数据的选样和未选样情况形成了鲜明对比。在保持识别率 100% 的前提下,本文样本选择算法最终所选出的训练样本个数不及原训练集样本数的 10% (分别为 9.3% 和 4.9%)。选样时间和测试时间合计也不及未经选样直接测试时间的 1% (实际分别仅为 0.59% 和 0.22%)。一方面,尽管本文算法在选样过程中也需要进行优化处理,但由于所涉及的数据规模较小,所以并不需要大量的时间处理,如两组实验的选样时间仅为 99s和 31s。另一方面,较少的选择样本也使得最近邻凸包分类器测试识别的优化代价大幅度降低,具体表现在测试所耗费的时间上。在两组实验中,未经选样的最近邻凸包分类器测试识别时间分别为 44 660s和 46 125s。与之相对照,选样后的分类器测试时间仅为 166s和 69s。显而易见,在该数据集上,本文的样本选择策略在保持最近邻凸包分类算法识别稳定的前提下,极大地缩短了系统的执行时间,提高了算法的执行效率。

在以上 ORL和 MIT-CBCL training-synthetic人脸数据库上的实验中,可以看出,随着选样数目的增加和逼近误差的减小,选择样本形成的凸包对训练集类别样本分布的估计能力不断提高,分类算法的泛化能力也得以不断增强,表现为整体测试识别率的稳步提高。而且该选样方法能够以较少的选择样本即可保证最近邻凸包分类器的泛化学习能力,在有效地降低分类算法优化计算代价的同时,使得最近邻凸包分类器在较大规模数据集上能够得以快速顺利运行。

## 5　结　论

最近邻凸包分类算法是一类将样本空间分布的凸包估计与最近邻分类思想相结合的分类方法。该方法以各类别所有训练样本作为相应类凸包生成元,以测试样本点到各类凸包的距离作为分类判别的依据。然而点到凸包距离的凸二次规划问题标准求解算法的较高计算复杂度限制了其在较大规模数据集上的应用。为此本文设计了一种样本选择方法。用各类别选出的较少的类凸包边缘点生成的子类凸包来近似各训练集类凸包,把子类凸包对样本点的逼近误差作为样本增选的依据。随着选择样本数目的增加,不断生长的子类凸包对训练样本的逼近误差不断减小,从而实现了对样本分布的有效估计。ORL人脸识别数据库和 MIT-CBCL人脸识别 training-synthetic库上的实验证实了这种方法的有效性。一方面本文方法选择的少量样本即可使最近邻凸包分类器达到较高的正确识别率,从而体现了选择样本凸包较强的类别表征能力。另一方面,较小的选样规模使得分类算法的计算代价大幅度的减少,在保证算法分类性能的同时,最近邻凸包分类器的分类速度较未经选样时的情况有了显著提高。

## 参考文献 (References)


1　Jiang Wen-Han, Zhou Xiao-fei, Yang Jing-yu. $p$-norm nearest neighbor convex hull classification algorithms[J]. Journal of Harbin Institute of Technology, 2006, 38(8): 982~984. [姜文瀚, 周晓飞, 杨静宇. $p$-范数最近邻凸包分类算法[J]. 哈尔滨工业大学学报, 2006, 38(8)(增刊): 982~984.]

2　Hyunjung Shin, Sungzoon Cho. Invariance of neighborhood relation under input space to feature space mapping[J]. Pattern Recognition Letters, 2005, 26: 707~718.

3　Bennett K, Bredensteiner E. Duality and geometry in SVM classifiers[A]. In: Proceedings of Seventeenth International Conference on Machine Learning[C], San Francisco, CA, USA: Morgan Kaufmann, 2000. 57~64.

4　Horst R, Pardalos P M, Thoai N V. Introduction to Global Optimization 2nd Edition[M]. Dordrecht, Netherland: Kluwer Academic Publishers, 2000.

5　The ORL Database of Face[DB/OL]. AT&T Laboratories Cambridge. http://www.cl.cam.ac.uk/Research/DTG/attarchive/facedatabase.html

6　Weyrauch B, Huang J, Heisele B, *et al*. Component-based face recognition with 3D morphable models[A]. First IEEE Workshop on Face Processing in Video[C], Washington, D.C., USA, 2004.