

基于 CSMDEM 算法的 GMM 学习方法

贾可新, 何子述

(电子科技大学电子工程学院, 成都 611731)

摘要: 基于 Mahalanobis 距离的 EM(MDEM)算法存在过分裂问题。为此, 提出一种竞争结束 MDEM(CSMDEM)算法。该算法将最小描述长度准则作为竞争结束条件嵌入到 MDEM 算法中, 能够在估计混合模型参数的同时选择模型阶数。实验结果表明, 该算法具有较低的平均 EM 迭代次数, 能够较好地拟合高斯混合模型。当其被应用到跳频网台分选时, 能够以较高的正确率分选跳频信号。

关键词: 高斯混合模型; Mahalanobis 距离; EM 算法; 最小描述长度准则

GMM Learning Method Based on CSMDEM Algorithm

JIA Ke-xin, HE Zi-shu

(School of Electronic Engineering, University of Electronic Science and Technology, Chengdu 611731, China)

[Abstract] To solve the over-splitting problem suffered in Mahalanobis distance based EM(MDEM) algorithm, a Competitive Stop MDEM (CSMDEM) algorithm is proposed. By regarding Minimum Description Length(MDL) criteria as a competitive stop condition and embedding it into MDEM algorithm, the CSMDEM algorithm can select model order while estimating the parameters of GMM. Experimental results show that the proposed CSEM algorithm has an increased capability to fit GMM while maintaining a low average number of EM iterations. By applying it to signal sorting, the proposed EM algorithm can sort FH signals with high correctness.

[Key words] Gaussian Mixture Model(GMM); Mahalanobis distance; Expectation Maximization(EM) algorithm; Minimum Description Length (MDL) criteria

DOI: 10.3969/j.issn.1000-3428.2011.19.050

1 概述

高斯混合模型已经被广泛地应用于数据统计建模的许多领域^[1-2], 如模式识别、计算机视觉、图像分析、复杂概率密度函数拟合等。在统计模式识别中, 与传统的基于启发式(如 K 均值)或层次聚类算法相比, 高斯混合模型为聚类分析提供了一种规范化的概率聚类方法。该方法认为高斯混合模型的混合分量是与数据中各类一一对应的, 单个观察样本是由某个高斯混合分量产生的, 各混合分量的参数估计实质上就是对数据进行聚类。用于拟合高斯混合模型的经典算法就是标准 EM 算法, 它能够给出混合模型参数的最大似然估计, 但标准 EM 算法存在 2 个缺陷^[3]: (1)模型阶数必须先已知; (2)对初始值敏感, 容易陷入局部极大值。为了解决标准 EM 算法中存在的问题, 许多改进的 EM 算法被提出。在文献[4]中, EM 算法起始于很高的模型阶数, 通过应用最小描述长度(Minimum Description Length, MDL)准则迭代删除混合分量, 直到找到最优模型阶数。而文献[5]提出了一种基于 Mahalanobis 距离的 EM(MDEM)算法, 它起始于单个混合分量, 采用基于 Mahalanobis 距离的正态性检验作为分裂准则, 算法终止于各混合分量都具有正态性。

本文在文献[5]的基础上, 将最小描述长度准则嵌入到 MDEM 算法, 提出了一种竞争结束 MDEM 算法, 有效地缓解了文献[5]算法的过分裂现象。

2 高斯混合模型与标准 EM 算法

2.1 高斯混合模型

设 $X=[X_1, X_2, \dots, X_D]^T$ 为 D 维随机向量, 而 $x=[x_1, x_2, \dots, x_D]^T$ 是随机向量 X 的一个观察样本。若随机向量 X 服从 Q 阶有限混合分布, 则其概率密度函数为:

$$p(x|\theta) = \sum_{m=1}^Q \alpha_m p(x|\theta_m) \quad (1)$$

其中, Q 是有限混合分布的阶数(混合分量的个数); $p(x|\theta_m)$ 是混合分布中第 m 个混合分量的概率密度函数; θ_m 是该分量的参数集; $\alpha_1, \alpha_2, \dots, \alpha_Q$ 是各个混合分量的混合概率; $\theta = \{\theta_1, \theta_2, \dots, \theta_Q, \alpha_1, \alpha_2, \dots, \alpha_Q\}$ 是描述该混合分布的完整参数集。对于混合概率 $\alpha_1, \alpha_2, \dots, \alpha_Q$, 它们应满足:

$$\alpha_m \geq 0, m=1, 2, \dots, Q, \sum_{m=1}^Q \alpha_m = 1 \quad (2)$$

假定上述 Q 阶有限混合分布中各混合分量的概率密度相同, 且都服从 D 维高斯分布, 则第 m 个混合分量的概率密度函数为:

$$p(x|\theta_m) = \frac{1}{(2\pi)^{D/2} |\Sigma_m|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_m)^T \Sigma_m^{-1}(x-\mu_m)\right) \quad (3)$$

其中, μ_m 是 D 维高斯随机矢量的均值矢量; Σ_m 是相应的协方差矩阵; $\theta_m = \{\mu_m, \Sigma_m\}$ 。因此, 随机向量 X 服从 K 阶高斯混合分布。

给定 N 个独立同分布的样本矢量集 $X = \{x_1, x_2, \dots, x_N\}$, 则 Q 阶高斯混合分布的对数似然函数为:

$$L(X|\theta) = \sum_{i=1}^N \ln \sum_{m=1}^Q \alpha_m p(x_i|\theta_m)$$

在实际中, 常采用式(1)给出的 Q 阶高斯混合分布拟合样本矢量集 X 概率密度函数, 即建立样本矢量集 X 的高斯混合模型, 这时需根据样本矢量集 X 确定 Q 阶高斯混合模型中各

作者简介: 贾可新(1982—), 男, 博士研究生, 研究方向: 通信信号分选, 模式识别; 何子述, 教授、博士生导师

收稿日期: 2011-03-22 **E-mail:** Jiakexin@sina.com

混合分量的参数 $\theta_m = \{\mu_m, \Sigma_m\}$ 和 α_m 。这些参数的最大似然估计(Maximum Likelihood Estimation, MLE)为:

$$\hat{\theta}_{MLE} = \arg \max_{\theta} \{L(X|\theta)\} = \arg \max_{\theta} \left\{ \sum_{i=1}^N \ln \sum_{m=1}^Q \alpha_m p(x_i|\theta_m) \right\} \quad (4)$$

2.2 标准 EM 算法

EM 算法将样本矢量集 $X = \{x_1, x_2, \dots, x_N\}$ 看作不完整数据, 该数据所丢失的部分是与样本矢量相对应的 N 个标签矢量集 $Z = \{z_1, z_2, \dots, z_N\}$ 。每一个标签矢量 $z_i = [z_1^{(i)}, z_2^{(i)}, \dots, z_k^{(i)}]$ 是一个二值矢量, 其中, $z_m^{(i)} = 1, z_p^{(i)} = 0, p \neq m$ 表示第 i 个样本矢量 x_i 是由高斯混合分布中第 m 个混合分量产生的。若完整的数据 $Y = (X, Z)$ 已经获得, 则完整的对数似然函数为:

$$\ln p(X, Z|\theta) = \sum_{i=1}^N \sum_{m=1}^Q z_m^{(i)} \ln [\alpha_m p(x_i|\theta_m)] \quad (5)$$

EM 算法通过交替应用 E-step 和 M-step 来获得参数估计序列 $\{\hat{\theta}(t), t = 0, 1, \dots\}$, 直到收敛。

E-step 给定当前参数估计值 $\hat{\theta}(t)$ 和样本矢量集 X , 计算完整对数似然函数 $\ln p(X, Z|\theta)$ 条件数学期望, 即 Q 函数:

$$Q(\theta, \hat{\theta}(t)) = E_Z \{ \ln p(X, Z|\theta) | X, \hat{\theta}(t) \} = \sum_{i=1}^N \sum_{m=1}^Q w_m^{(i)} \ln [\alpha_m p(x_i|\theta_m)] \quad (6)$$

其中:

$$w_m^{(i)} = \Pr \{ z_m^{(i)} = 1 | X, \hat{\theta}(t) \} = \frac{\hat{\alpha}_m(t) p(x_i|\hat{\theta}_m(t))}{\sum_{j=1}^Q \hat{\alpha}_j(t) p(x_i|\hat{\theta}_j(t))} \quad (7)$$

M-step 在 $\hat{\theta}(t)$ 已知的情况下, 通过求 $Q(\theta, \hat{\theta}(t))$ 关于 θ 的最大值, 可得高斯混合模型的参数更新公式为:

$$\hat{\mu}_m(t+1) = \frac{\sum_{i=1}^N w_m^{(i)} x_i}{\sum_{i=1}^N w_m^{(i)}} \quad (8)$$

$$\hat{\Sigma}_m(t+1) = \frac{\sum_{i=1}^N w_m^{(i)} (x_i - \hat{\mu}_m(t+1))(x_i - \hat{\mu}_m(t+1))^T}{\sum_{i=1}^N w_m^{(i)}} \quad (9)$$

$$\alpha_m(t+1) = \frac{1}{N} \sum_{i=1}^N w_m^{(i)} \quad (10)$$

3 竞争结束 MDEM 算法

本节将样本矢量集 X 建模为服从高斯混合分布的总体的一组随机样本矢量。也就是说, X 可以被看作是 Q 个类 $L_q (q=1, 2, \dots, Q)$ 的并集, 且每一个类 L_q 中的样本矢量是服从多维高斯分布的总体的一组随机样本矢量。本文提出算法的目标就是寻找 $\{L_q\}_{q=1}^Q$, 算法的流程如图 1 所示。

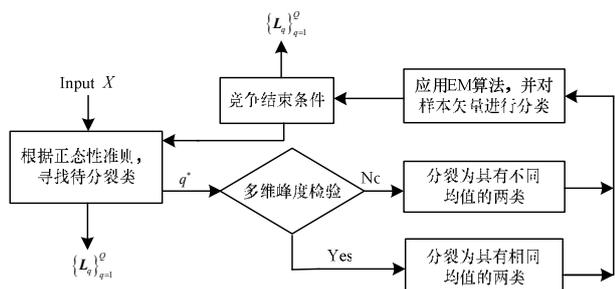


图 1 本文算法流程

3.1 待分裂类

本节以基于 Mahalanobis 距离的多维正态性检验作为分裂准则。为了寻找待分裂类, 根据文献[5]中正态性检验方法对 $\{L_q\}_{q=1}^Q$ 进行检验, 可得各类的检测统计量 $\{D_{L_q}\}_{q=1}^Q$, 则对于所有 $q=1, 2, \dots, Q$, 若有:

$$D_{L_q} < (1-\lambda) |L_q| \quad (11)$$

成立, 则算法终止, 因为没有类偏离高斯分布。否则, 从选择类 L_q 作为待分裂类, 如果它满足:

$$q^* = \arg \max_{q=1, 2, \dots, Q} \{D_{L_q} - (1-\lambda) |L_q|\}$$

其中, $|L_q|$ 是第 q 类的容量(样本矢量个数)。

3.2 分裂操作

设 $L_q = \{x_{q'i}\}_{i=1}^{N_q}$ 为待分裂类, 本节根据文献[5]中的多维峰度检验方法将 L_q 分裂为 L'_q 和 L'_Q , 其中, $Q' = Q+1$ 。

计算 $K(L'_q)$ 和 K_0 。若 $K(L'_q) > K_0$, 则将 L'_q 和 L'_Q 的均值初始化为 $\mu'_q = \mu_q, \mu'_Q = \mu_q$ 。

混合概率初始化为 $\alpha'_q = \alpha_q/2, \alpha'_Q = \alpha_q/2$, 而新类的协方差矩阵 Σ'_q, Σ'_Q 按如下方式进行随机初始化: 设置 Σ'_q, Σ'_Q 等于 2 个不同的 $D \times D$ 维对角矩阵, 每个矩阵的对角元素为随机变量 σ^2 的样本。

随机变量 σ^2 满足:

$$\sigma^2 \frac{2D(|L_q| - 1)}{\|\Sigma_q\|} \square \chi^2_{|L_q|-1} \quad (12)$$

其中, $\|\Sigma_q\|$ 为 Σ_q 的行列式。协方差矩阵的随机初始化是基于这样一个事实: 边缘方差应该服从自由度为 $|L_q|-1$ 的 χ^2 分布。

若 $K(L'_q) < K_0$, 则根据判别超平面将 L_q 分裂为 L'_q 和 L'_Q 。该判别超平面由样本值 $x_{q'i}^{(d)}$ 决定, 即:

$$x_{q'i}^{(d)} = \arg \max_{\substack{i=1, 2, \dots, |L_q| \\ d=1, 2, \dots, D}} \left\{ F_{\chi^2_{|L_q|}}(x_{q'i}^{(d)}) - \hat{F}_{\chi^2_{|L_q|}}(x_{q'i}^{(d)}) \right\} \quad (13)$$

由样本值 $x_{q'i}^{(d)}$ 决定的超平面与 $X_q^{(d)}$ 轴垂直, 它能够将 1 个类分裂为 2 个均值不同的类。当以上分裂操作结束以后, 2 个新类的均值、协方差矩阵和先验概率用于初始化 EM 算法。

3.3 EM 算法应用

为更新高斯混合模型的参数, 将 EM 算法的初始参数值 $\theta(0) = \{\alpha_q(0), \mu_q(0), \Sigma_q(0)\}_{q=1}^Q$ 设置为:

$$\alpha_q(0) = \frac{|L_q|}{|X|}$$

$$\mu_q(0) = \frac{1}{|L_q|} \sum_{x \in L_q} x$$

$$\Sigma_q(0) = \frac{1}{|L_q|-1} \sum_{x \in L_q} (x - \mu_q)(x - \mu_q)^T$$

其中, $q=1, 2, \dots, Q'$; $|L_q|$ 为 L_q 的容量。

EM 算法的收敛条件为:

$$|L(X|\theta(t+1)) - L(X|\theta(t))| < 10^{-5} |L(X|\theta(t))| \quad (14)$$

若满足上式, 则由 EM 算法可获得各样本矢量属于 $\{L_q\}_{q=1}^Q$ 的后验概率。此时可根据最大后验准则对各样本矢量重新进行分类。

3.4 竞争结束条件

为了防止 EM 算法过度分裂, 本文引入了一个竞争结束条件。当满足该条件时, 算法也将终止。竞争结束条件是以 MDL 准则^[3-4]:

$$L_2(\theta) = \frac{1}{2} \left[Q-1 + Q \left(D + \frac{D(D+1)}{2} \right) \right] \ln(N) - \ln p(X|\theta) \quad (15)$$

为基础的。设混合分量的个数为 Q 时的 MDL 值为 $L_2^{(Q)}$, 若有:

$$L_2^{(Q)} \leq L_2^{(Q+1)} \quad (16)$$

成立, 则 EM 算法结束。此时 EM 算法输出 $L_2^{(Q)}$ 对应的分裂结果。

EM 算法从 3.1 节~3.4 节重复执行, 直到所有的类都具有正态性或满足竞争结束条件。此时, 将获得最终的聚类结果。

4 仿真实验

为了验证本文所提出算法的有效性, 本节首先将竞争结束 MDEM 算法与文献[5]提出的 MDEM 算法进行比较, 然后给出竞争结束 MDEM 算法的一个新的应用-跳频信号分选。

4.1 竞争结束 MDEM 算法的性能

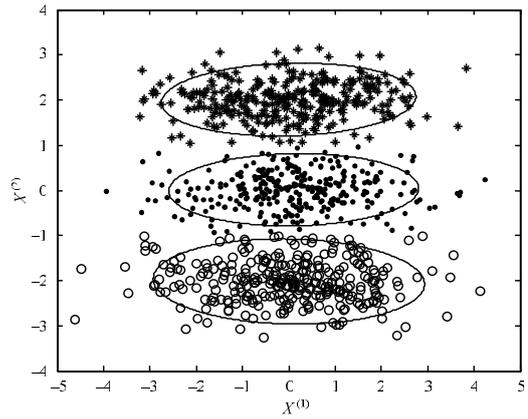
本节采用文献[5]中图 13 给出的 3 组仿真数据对竞争结束 MDEM 算法的性能进行测试。这 3 组数据在改进的 EM 算法研究中, 被广泛地用于验证算法的聚类性能。评价算法性能的指标包括类别正确识别率和平均 EM 迭代次数。仿真实验环境为 Intel Celeron 2.52 GHz CPU 和 512 MB RAM Matlab7.6。

对 3 组仿真数据进行 1 000 次蒙特卡罗实验, 本文提出算法测试结果如表 1 所示。MDEM 算法与其他 7 种改进的 EM 算法的测试结果在文献[5]的表 2 中已经列出。对比表 1 和文献[5]中的表 2 可知, 由于竞争结束条件的引入, 因此本文所提出算法在保持较低的平均 EM 迭代次数的同时, 对 3 组测试数据具有最高的聚类正确率。与 MDEM 算法相比, 对于测试数据 A 和数据 C, 本文提出的算法类别正确率都为 100%, 而平均迭代次数分别降低了 4 倍和 2.3 倍。对于测试数据 C, 类别正确识别率也有明显的提高, 但平均迭代次数的改善并不明显。

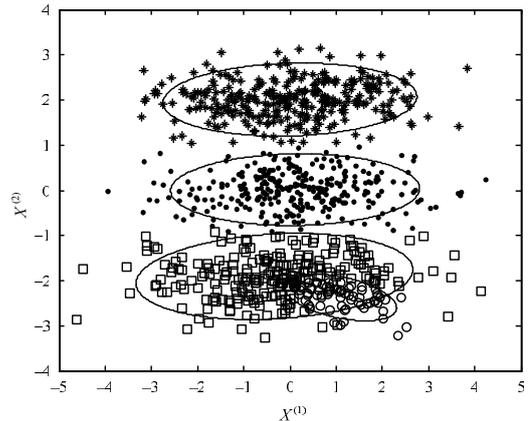
表 1 各样本集的测试结果

| 算法 | 测试数据 | 类别正确率/(%) | 平均 EM 迭代次数 |
|--------|------|-----------|------------|
| CSMDEM | A | 100.0 | 20 |
| | B | 84.1 | 96 |
| | C | 100.0 | 115 |

图 2 给出了在对测试数据 A 进行聚类时, 满足竞争结束条件的一个实例。测试数据 A 的一组随机样本经第 2 次分裂后, 其聚类结果如图 2(a)所示, 此时 MDL 值 $L_2^{(3)} = 3 126.0$ 。由于原始 MDEM 算法出现过分裂现象, 因此 EM 算法并没有在第 2 次分裂后终止, 而是执行了第 3 次分裂, 其聚类结果如图 2(b)所示, 此时 MDL 值 $L_2^{(4)} = 3 142.7$, 满足竞争结束条件, 竞争结束 EM 算法终止, 并以图 2(a)作为最终聚类结果。



(a)第 2 次分裂后聚类结果



(b)第 3 次分裂后聚类结果

图 2 满足竞争结束条件的实例

4.2 跳频信号分选

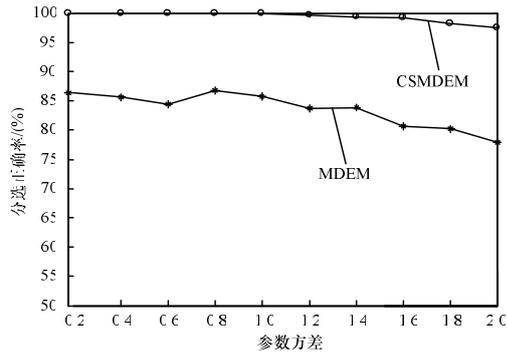
本节采用 2 组不同的特征参数验证竞争结束 EM 算法在跳频信号分选中的性能, 仿真条件如表 2 所示。表 2 中 δ 表示特征参数的方差, 它在 [0.2, 2.0] 之间取值。第 1 组特征参数仅包括方位角和俯仰角, 它可用于跳频通信中同步正交组网电台的网内分选。因同步网^[6]内的各电台同时改变载波, 具有相同的跳时和跳周期, 但各电台的波达方向(俯仰角和方位角)一般不会相同。第 2 组特征参数包括方位角、俯仰角和跳时, 它可用于跳频通信中异步非正交组网电台的网内分选。因异步非正交网^[6]内的各电台使用相同的跳频频率序列, 一般具有相同的跳周期, 但各电台的波达方向(俯仰角和方位角)和跳时不会相同。

表 2 跳频信号分选的仿真参数

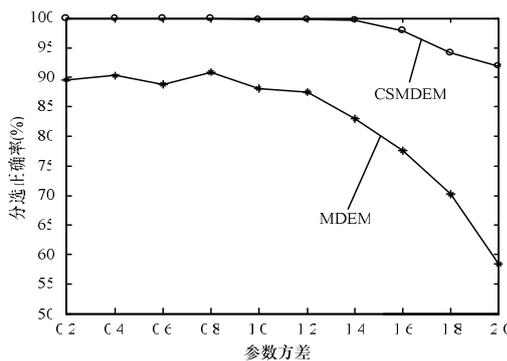
| 组号 | 特征参数 | 仿真参数 |
|----|--------|--|
| 1 | 方位角(°) | $N=1 200, Q=4, \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.25$ $\mu = \begin{bmatrix} 30 & 35 & 40 & 45 \\ 50 & 53 & 55 & 60 \end{bmatrix}, \Sigma = \begin{bmatrix} \delta & 0 \\ 0 & \delta \end{bmatrix}$ |
| | 俯仰角(°) | |
| 2 | 方位角(°) | $N=900, Q=3, \alpha_1 = \alpha_2 = \alpha_3 = 1/3$ $\mu = \begin{bmatrix} 70 & 74 & 78 \\ 50 & 53 & 55 \\ 12 & 10 & 8 \end{bmatrix}, \Sigma = \begin{bmatrix} \delta & 0 & 0 \\ 0 & \delta & 0 \\ 0 & 0 & \delta \end{bmatrix}$ |
| | 俯仰角(°) | |
| | 跳时/ms | |

图 3(a)和图 3(b)分别给出了上述 2 组数据的分选正确率随着特征参数方差 δ 的变化曲线。对于 MDEM 算法, 受过分裂现象的影响, 即使在参数方差很小的情况下, 它也不能以很高的正确率分选跳频信号。而 CSMDEM 算法在方差小于 1.4 时, 分选正确率在 99% 以上。本文提出的 CSMDEM 算法

具有更高的分选正确率。



(a) 第1组特征参数的分选正确率



(b) 第2组特征参数的分选正确率

图3 跳频信号分选正确率

5 结束语

为了更加有效地拟合高斯混合模型,解决 MDEM 算法中

存在的过分裂问题,本文提出了一种竞争结束 MDEM 算法。该算法以基于 Mahalanobis 距离的正态性检验作为分裂准则,而以所有类都具有正态性或满足竞争结束条件作为分裂终止条件。仿真实验结果表明,基于 MDL 准则的竞争结束条件的引入有效地解决了文献[5]中存在的过分裂问题。当将 CSMDDEM 算法应用于跳频信号分选时,它能够准确地分选跳频信号。值得注意的是,本文虽然仅采用方位角、俯仰角和跳时进行跳频信号分选,但是该方法可以很容易地推广到基于其他特征参数的跳频信号分选。

参考文献

- [1] 程远国, 耿伯英. 基于高斯混合模型的无线局域网定位算法[J]. 计算机工程, 2009, 35(4): 25-27.
- [2] 许晓东, 熊卫斌, 朱士瑞. 基于高斯混合模型的流量矩阵估算研究[J]. 计算机工程, 2009, 35(14): 132-134.
- [3] McLachlan G, Peel D. Finite Mixture Models[M]. New York, USA: John Wiley & Sons, 2000.
- [4] Figueiredo M A, Jain A K. Unsupervised Learning of Finite Mixture Models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(3): 381-396.
- [5] Ververidis D, Kotropoulos C. Gaussian Mixture Modeling by Exploiting the Mahalanobis Distance[J]. IEEE Transactions on Signal Processing, 2008, 56(7): 2797-2811.
- [6] 梅文华, 王淑波, 邱永红, 等. 跳频通信[M]. 北京: 国防工业出版社, 2005.

编辑 索书志

(上接第 147 页)

计算量较大的原始特征, 本文融合策略在速度上的优势会更加明显。

6 结束语

本文提出了一种基于全局和局部特征融合的人脸识别方法。针对实际应用中对面脸识别系统速度和精度的不同要求, 给出 2 种融合策略组合全局和局部特征。在 FRGC v2.0 大规模人脸库上的实验结果表明: (1) 本文的多特征“局部+全局”分数层融合模式可以有效地提高系统整体性能; (2) 在对特征长度、计算时间以及内存有比较严格的限制, 导致可能无法进行多特征全局特征融合的情况下, 以某一种原始特征的全局特征与其他特征的局部特征进行融合的模式可以以相对“经济”的方式达到较好的系统性能。

本文仅以空间位置来划分全局和局部特征的方式还显得比较粗糙, 下一步工作还需要对人脸全局、局部特征的划分方式进行扩展, 进一步研究人类视觉感知系统中全局和局部特征相互作用的机理, 从而更好地利用计算机系统区分来自不同个体的人脸图像。

参考文献

- [1] Chellapa R, Wilson C L, Sirohey S. Human and Machine Recognition of Faces: A Survey[J]. Proceedings of the IEEE, 1995, 83(5): 705-740.

- [2] Daugman J G. Uncertainly Relation for Resolution in Space, Spatial Frequency, and Orientation Optimized by Two-dimensional Visual Cortical Filters[J]. Journal of the Optical Society of America A, 1985, 2(7): 1160-1169.
- [3] Ahonen T, Hadid A, Pietikainen M. Face Description with Local Binary Patterns: Application to Face Recognition[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2006, 28(12): 2037-2041.
- [4] Dalal N, Triggs B. Histogram of Oriented Gradients for Human Detection[C]//Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition. San Diego, USA: [s. n.], 2005: 886-893.
- [5] Shu Chang, Ding Xiaoqing, Fang Chi. Histogram of the Oriented Gradient for Face Recognition[J]. Tsinghua Science and Technology, 2011, 16(2): 216-224.
- [6] 叶剑华, 刘正光. 基于 LBP 和 Fisherfaces 的多模态人脸识别[J]. 计算机工程, 2009, 35(11): 193-195.
- [7] Phillips P J, Flynn P J, Scruggs T, et al. Overview of the Face Recognition Grand Challenge[C]//Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition. San Diego, USA: [s. n.], 2005: 947-954.

编辑 索书志