

一种序列的加权 kNN 分类方法

朱明早^{1,2}, 罗大庸¹, 易励群³

(1. 中南大学信息科学与工程学院, 湖南长沙 410083; 2. 湖南文理学院电气与信息工程学院, 湖南常德 415000;
3. 湖南文理学院现代教育技术中心, 湖南常德 415000)

摘 要: 针对加权 kNN (k -Nearest Neighbor) 方法在对样本进行分类时, 仅仅只利用了它的 k 近邻点来进行分类决策的不足, 提出了一种序列的加权 kNN 分类方法. 该方法在对某个测试样本进行分类时, 除了利用它 k 近邻点所提供的类别信息外, 还有效地利用了前面已分类样本的类别信息, 这使得测试样本的分类决策更加合理和有效. 在 Cohr Kanade 人脸库上进行的表情识别实验表明, 在序列样本分类的场合, 该方法的分类效果比加权 kNN 方法更好.

关键词: 加权 kNN; 流形; 贝叶斯规则; 序列的加权 kNN

中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2009) 11-2584-05

A Sequential Weighted k -Nearest Neighbor Classification Method

ZHU Ming-han^{1,2}, LUO Da-yong¹, YI Li-qun³

(1. College of Information Science and Engineering, Central South University, Changsha, Hunan, 410083, China;
2. College of Communication and Electric Engineering, Hunan University of Arts and Science, Changde, Hunan 415000, China;
3. Modern Education Technology Center, Hunan University of Arts and Science, Changde, Hunan 415000, China)

Abstract: Aim at the defect that weighted k -nearest neighbor method classifies one test sample only using the class information of its k -nearest samples, a sequential weighted k -nearest neighbor classification method is proposed in this paper. Not only the class information offered by k -nearest neighbor points of test sample but also the class information of previous test sample is used for classification in the proposed method. So its decision-making processing is more reasonable and effective. The experimental results of facial expression recognition in Cohr Kanade face database show the method is better than weighted k -nearest neighbor method for the classification of sequential samples.

Key words: weighted k -nearest neighbor; manifold; Bayesian rule; sequential weighted k -nearest neighbor

1 引言

kNN (k -Nearest Neighbor) 分类方法是最近邻方法^[1]的一个推广, 它将测试样本分类为与它最接近的 k 个近邻中出现最多的那个类别. 这种非参数分类方法, 对许多种数据集的分类都十分有效^[2,3]. 已被广泛应用于分类、回归和模式识别等领域中.

k 近邻方法在决定测试样本的类别时, 赋予了这 k 个近邻样本平等的贡献. 根据贝叶斯决策规则, 为了得到可靠的分类结果, 一方面要使 k 越大越好, 另一方面又要使测试样本的 k 个近邻样本距它越近越好. 因此, 常需根据实际情况, 在选取 k 值时, 做出某种折中^[4]. 对此, 一些学者提出了一些最优 k 值的搜索方案, 如: 先给 k 设定一个初值, 然后不断地调整 k 值, 并用留一法 (Leave-One-Out, LOO) 进行实验, 根据实验结果得到最优 k 值, 这种方法用起来相当费时. 后来, Gra 等人^[5]提出

了一种自动选取最优 k 值的 k 近邻方法. Hechenb-ichler 等人^[6]则提出了一种加权 kNN 方法, 该方法根据各近邻样本到测试样本的距离 (或其它相似度) 的大小, 赋予这 k 个近邻样本不同的权值. 距离越小权值就越大, 相反, 距离越大, 所赋予的权值就越小. 这样, 近邻样本与测试样本的相似程度, 就通过权值的大小来体现. 因此, 既使 k 值很大, 对测试样本分类起决定作用的, 仍是与它相距较近的那些样本. 这种加权 k 近邻方法的分类准确率, 对 k 值的选取不再敏感, 表现出了较好的鲁棒性.

另一些学者则对相似度的确定问题进行了研究, 提出在计算近邻样本与测试样本的相似程度时, 对它们的各个特征进行加权. 相关性越强, 赋予的权值越大, 相关性越弱, 赋予的权值越小, 对不相关的特征, 则赋予 0 权值. 如: 陈振洲^[7]等人提出的基于 SVM 的特征加权 kNN 算法; 刘明^[8]等人提出的证据理论 k 近邻规则中, 相似

收稿日期: 2008-09-18; 修回日期: 2009-06-18

基金项目: 湖南省教育厅科研项目 (No. 08C606); 国家自然科学基金项目 (No. 60776834)

度参数的确定方法; Vivencio 等人^[9]提出的基于 χ^2 统计检验的特征加权 kNN 算法, 以及孙岩^[10]等人提出的基于贝叶斯结构特征加权的 kNN 算法. 这些算法加重了相关特征对分类所起的作用, 在一定程度上提高了分类的准确率.

然而, 以上这些加权 kNN 算法, 在对样本进行分类时, 都只利用了待测样本的 k 近邻点所提供的类别信息, 而没有考虑已测样本与该待测样本的内

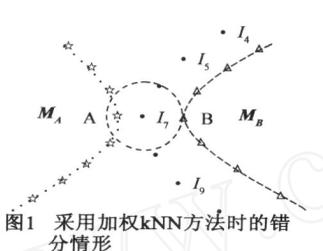


图1 采用加权kNN方法时的错分情形

在联系. 若用于对序列中的样本进行分类, 就会导致一些明显的误判. 例如: 在人脸图像序列中的人脸识别; 在表情图像序列中的表情识别等等. 我们用图 1 展示了用加权 k 近邻方法对序列样本进行分类时, 出现的一些错分现象. 图中 M_A 和 M_B 分别表示两个不同的流形, I_4, \dots, I_{10} 是一测试样本序列, 若用加权 k 近邻方法, I_7 会被归为 A 类. 实际上, 综观整个测试样本序列, 不难发现 I_7 同其它的测试样本一样, 也属于 B 类.

针对这个缺陷, 本文从加权 k 近邻方法的基本原理入手, 提出了一种序列的加权 kNN 分类方法, 该方法在对序列 $I_0, I_1, \dots, I_t, \dots$ 中的样本 I_t 进行分类时, 不仅利用了它 k 近邻点所提供的类别信息, 而且还考虑了已测样本 I_0, \dots, I_{t-1} 对 I_t 分类的影响, 使分类的信息更加全面、可靠, 从而大大提高了序列样本分类的准确率.

2 加权 kNN 原理

设 $L = \{(y_i, x_i), x_i \in R^d, i = 1, 2, \dots, n\}$ 是一个由 n 个样本组成的训练集, 每个样本 x_i 的类别标示 y_i 均已知, $y_i \in \{1, \dots, r\}$. x_t 为待测样本, 它的类别 y_t 待测. 加权 kNN 分类方法的原理如下:

(1) 根据明考斯基(Minkowski) 距离

$$d(x_t, x_i) = \left(\sum_{j=1}^d |x_{tj} - x_{ij}|^q \right)^{1/q}$$
 (或其它特征加权相似度), 从训练集 L 中, 找出 x_t 的 $k+1$ 个近邻样本. 用 $x_{t(1)}, x_{t(2)}, \dots, x_{t(k+1)}$ 表示这 $k+1$ 个近邻样本, $y_{t(1)}, y_{t(2)}, \dots, y_{t(k+1)}$ 表示它们的类别标记.

(2) 用第 $k+1$ 个近邻样本到 x_t 的距离 $d(x_t, x_{t(k+1)})$, 标准化前 k 个近邻样本到 x_t 的距离:

$$D(x_t, x_{t(l)}) = \frac{d(x_t, x_{t(l)})}{d(x_t, x_{t(k+1)})}, \quad l = 1, 2, \dots, k \quad (1)$$

(3) 用加权核函数 $K(\cdot)$, 将标准距离 $D(x_t, x_{t(l)})$ 转化为 x_t 与 $x_{t(l)}$ 同类的概率 $P(x_{t(l)} | x_t)$. 若 $K(\cdot)$ 为高斯核函数则有:

$$P(x_{t(l)} | x_t) = \frac{1}{\sqrt{2}} \exp\left(-\frac{D(x_t, x_{t(l)})}{2}\right) \quad (2)$$

常用的加权核函数 $K(\cdot)$ 有八种, 详细内容可参考文献 [6].

(4) 根据 x_t 的这 k 个近邻样本, 求出 x_t 为类别 s ($s = 1, 2, \dots, r$) 的后验概率

$$P(s | x_t) = \frac{1}{\sum_{l=1}^k P(x_{t(l)} | x_t) I(y_{t(l)} = s)} \quad (3)$$

$$I(A) = \begin{cases} 1, & \text{如果 } A \text{ 为真} \\ 0, & \text{如果 } A \text{ 为假} \end{cases}$$

这里 $\frac{1}{\sum_{l=1}^k P(x_{t(l)} | x_t)}$ 为归一化因子. 最大 $P(s | x_t)$ 所对应的类别被判定为 x_t 的类别, 即:

$$t = \arg \max_s P(s | x_t) \quad (4)$$

3 序列的加权 kNN

设 $T = \{x_0, x_1, \dots, x_t\}$ 为一样本序列, 考虑对序列 T 的各个样本进行分类. 若每次分类都采用上述的加权 kNN 方法, 那么各测试样本间的序列联系没有得到任何利用, 这显然具有不合理的一面.

考虑到 x_t 与 $x_{0:t-1}$ 的序列关系, 在对样本 x_t 分类时, 同文献 [11] 一样, 我们采用了联合后验概率 $P(s | x_t, x_{0:t-1})$ 作为它的分类依据. 根据式 (3) 有

$$P(s | x_t, x_{0:t-1}) = \frac{1}{\sum_{l=1}^k P(x_{t(l)} | x_{0:t})} P(x_{t(l)} | x_{0:t}) I(y_{t(l)} = s)$$

这里 $x_{0:t}$ 表示序列 x_0, x_1, \dots, x_t . 假设 x_t 与 $x_{0:t-1}$ 在给定的 $x_{t(l)}$ 下是条件独立的, 运用贝叶斯理论^[12], 有如下的推理:

$$P(x_{t(l)} | x_{0:t}) = \frac{1}{\prod_1} P(x_t | x_{t(l)}, x_{0:t-1}) P(x_{t(l)} | x_{0:t-1})$$

$$= \frac{1}{\prod_1} P(x_t | x_{t(l)}) P(x_{t(l)} | x_{0:t-1}) \quad (5)$$

这里 $\prod_1 = P(x_t | x_{0:t-1})$. 同理, 对式 (5) 继续递推

$$P(x_{t(l)} | x_{0:t}) = \frac{1}{\prod_{1,2}} P(x_t | x_{t(l)}) P(x_{t-1} | x_{t(l)}) P(x_{t(l)} | x_{0:t-2})$$

$$\dots$$

$$= \frac{1}{\prod_{j=1}^t} P(x_j | x_{t(l)}) P(x_{t(l)} | x_0)$$

这里, $\prod_j = P(x_{t+1-j} | x_{0:t-j}), 1 \leq j \leq t, \prod_1 = \prod_{j=1}^t$. 因此, 联合后验概率为

$$P(s | x_{0:t}) = \frac{\prod_{l=1}^k \prod_{j=1}^t P(x_j | x_{t(l)}) P(x_{t(l)} | x_0) I(y_{t(l)} = s)}{\prod_{l=1}^k \prod_{j=1}^t P(x_j | x_{t(l)}) P(x_{t(l)} | x_0)}$$

$$= \frac{1}{\sum_{l=1}^k \prod_{j=1}^t} P(x_j | x_{t(l)}) P(x_{t(l)} | x_0) I(y_{t(l)} = s) \quad (6)$$

$$* = \prod_{l=1}^k \prod_{j=1}^t P(x_j | x_{t(l)}) P(x_{t(l)} | x_0)$$

根据加权 kNN 原理, x_t 与 $x_{t(l)}$ 同类的概率 $P(x_0 | x_{t(l)}) = P(x_{t(l)} | x_0)$, 另外, 式(6)中, 对不同的 s , $*$ 的值都相等, 为了简化计算, 我们用下式计算 x_t 属于类别 t 的度量值:

$$P^*(s | x_{0:t}) = \prod_{l=1}^k \prod_{j=0}^t P(x_{t(l)} | x_j) I(y_{t(l)} = s) \quad (7)$$

根据上述的结论, 有如下的序列加权 kNN 分类方法:

设 $L = \{(y_i, x_i), x_i \in R^d, i = 1, 2, \dots, n\}$ 是一个由 n 个样本组成的训练集, 每个样本 x_i 的类别标示 y_i 均为已知, $y_i \in \{1, \dots, r\}$. x_t 为一待测样本, 它来自于序列 $\{x_0, x_1, \dots, x_t, \dots\}$, 它的类别 t 待测.

(1) 根据明考斯基距离 $d(x_t, x_i) = (\sum_{j=1}^d |x_{tj} - x_{ij}|^q)^{1/q}$ 或其它特征加权距离, 从训练集 L 中, 找出 x_t 的 $k+1$ 个近邻点.

(2) 用第 $k+1$ 个近邻点到 x_t 的距离 $d(x_t, x_{t(k+1)})$, 标准化 k 近邻点到 $\{x_j, j=0, 1, \dots, t\}$ 的距离.

$$D_{j(l)} = \frac{d(x_t, x_{t(l)})}{d(x_t, x_{t(k+1)})}, j=0, 1, \dots, t; l=1, 2, \dots, k \quad (8)$$

(3) 用高斯核函数, 将标准距离 $D_{j(l)}$ 转化为 x_j 与 $x_{t(l)}$ 同类的概率 $P(x_{t(l)} | x_j)$.

$$P(x_{t(l)} | x_j) = \frac{1}{\sqrt{2}} \exp(-\frac{D_{j(l)}}{2}) \quad (9)$$

(4) 用式(10) 求出样本 x_t 属于类别 t 的概率 $P^*(s | x_{0:t}) (s=1, 2, \dots, r)$.

$$P^*(s | x_{0:t}) = \prod_{l=1}^k \prod_{j=0}^t P(x_{t(l)} | x_{t-j}) I(y_{t(l)} = s) \quad (10)$$

最大 $P^*(s | x_{0:t})$ 值所对应的类别就是 x_t 的类别 t , 即:

$$t = \arg \max_{l=1}^k \prod_{j=0}^t P(x_{t(l)} | x_{t-j}) I(y_{t(l)} = s)$$

在实际应用中, 可以只估计 x_t 与它前 c 个样本 (c 为一固定常数) 的联合后验概率, 并将其作为它的分类依据, 即:

$$t = \arg \max_{l=1}^k \prod_{j=0}^c P(x_{t(l)} | x_{t-j}) I(y_{t(l)} = s) \quad (11)$$

假设在 d 维空间中, 有 n 个已标记的训练样本. 在最简单的方法中, 计算每一个距离的计算复杂度为 $O(d)$, 搜索方法的总计算复杂度为 $O(dn^2)$. 将距离转化为概率, 并求出待测样本属于各个类别概率的计算复杂度为 $O(1)$. 所以加权 kNN 方法, 总的计算复杂度

为 $T = O(dn^2) + O(1) = O(dn^2)$. 序列加权 kNN 方法的第一个步骤与加权 kNN 方法一样, 而后续的三个步骤均与训练样本的维数 d 和规模 n 无关, 计算复杂度也为 $O(1)$. 可见, 这两种算法具有相同的计算复杂度.

为了比较直观地说明序列加权 kNN 方法的优势, 下面我们分别用两种方法来对图 1 中的待测样本 I_7 进行分类. 为了阐述的方便, 取 $k=2, c=2$, 则 A 和 B 分别为 I_7 的两个最近邻样本. 不难发现

$$d(I_7, A) < d(I_7, B), d(I_6, A) > d(I_6, B), \\ d(I_5, A) > d(I_5, B).$$

根据式(2), 有

$$P(A | I_7) > P(B | I_7), P(A | I_6) < P(B | I_6), \\ P(A | I_5) < P(B | I_5).$$

不妨假设

$$P(A | I_7) = 0.8, P(B | I_7) = 0.6, P(A | I_6) = 0.38, \\ P(B | I_6) = 0.4, P(A | I_5) = 0.1, P(B | I_5) = 0.2.$$

用加权 kNN 方法, 得到 I_7 属于 A 类和 B 类的概率分别为

$$P(A | I_7) = 0.8 / 1.4 = 0.57 \\ P(B | I_7) = 0.6 / 1.4 = 0.43$$

$P(A | I_7) > P(B | I_7)$, I_7 被地判定为 A 类, 分类出现了错误. 用本文的方法, 根据式(11) 求得 I_7 属于 A 和 B 类的度量值为

$$P^*(A | I_{5:7}) = 0.8 \times 0.38 \times 0.1 = 0.0304 \\ P^*(B | I_{5:7}) = 0.6 \times 0.4 \times 0.2 = 0.0480$$

$P^*(A | I_{5:7}) < P^*(B | I_{5:7})$, I_7 被判定为 B 类, 分类正确. 由于序列的加权 kNN 方法利用了 I_6 和 I_5 提供的类别信息, 因而提高了分类的准确性.

4 实验与分析

4.1 实验

实验是在 Cohn Kanade 人脸库^[13] 上进行的, Cohn Kanade 人脸库含有 210 个对象的六种表情 (高兴、悲伤、惊讶、生气、厌恶、恐惧) 序列. 我们从 Cohn Kanade 人脸库中选取了 15 个对象的六种表情序列, 每个表情序列选 12 幅图像, 共 1080 幅图像, 图 2 为 Cohn Kanade 人脸库中的一些样本.



图2 Cohn-Kanade人脸库中的样本

我们先用等距映射 (Isomap) 算法^[14], 将原始图像投影到 7 维流形空间. 图 3(a) 是 15 个人的图像经等距映射后得到的表情流形, (b) 为 (a) 图中某一人的表情流

形的放大图(为了可视化我们只取了流形坐标的前3个值作为 x, y, z 的坐标值,图中 * 代表厌恶, 代表惊奇, 代表悲伤, 代表生气, 代表恐惧, + 代表高兴).

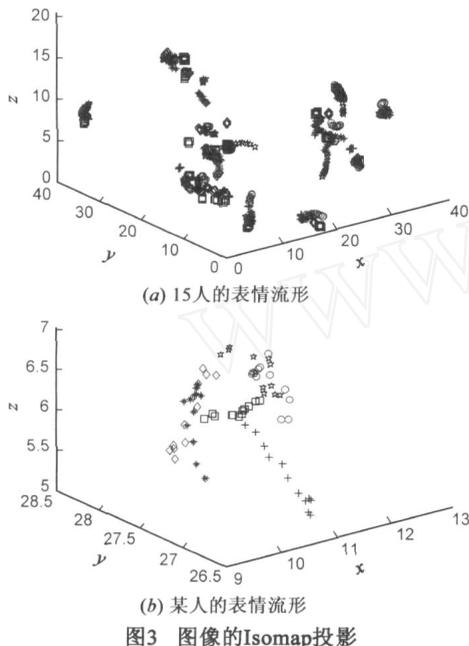


图3 图像的Isomap投影

然后,将这 1080 个低维样本分为两个样本集 X_1 和 X_2 ,各由 15 人的 6 种表情序列组成,每种表情序列包含 6 个不同强度的表情样本.我们采用交叉验证的方法(Cross Validation)进行了两种表情识别实验:实验 1)测试样本是从测试集的 540 个样本中随机取出,前后测试样本间没有序列关系;实验 2)测试样本以表情序列为组进行识别,共有 90(6 × 15) 个表情序列,每组的各测试样本间存在序列关系.每种实验分别采用了加权 kNN 与本文提出的序列的加权 kNN 方法来进行表情分类. $k = 5$, k 近邻点的判断采用了欧氏距离 $q = 2$,实验的识别率数据如表 1 和表 2 所示.

表 1 实验 1 的表情识别率数据

	高兴	悲伤	厌恶	惊讶	生气	恐惧
加权 kNN	89.4 %	81.2	82.3 %	85.5 %	87.2 %	82.7 %
本文的方法	91.2 %	82.6 %	83.4 %	85.2 %	87.2 %	82.9 %

表 2 实验 2 的表情识别率数据

	高兴	悲伤	厌恶	惊讶	生气	恐惧
加权 kNN	89.4 %	81.2	82.3 %	85.5 %	87.2 %	82.7 %
本文的方法	95.2 %	90.6 %	90.4 %	93.1 %	94.2 %	91.6 %

4.2 实验分析

综合表 1 的数据我们得到,用加权 kNN 方法得到的平均识别率为 84.72 %,本文方法得到的为 85.42 %.这说明在前后测试样本间不存在序列关系的情况下,两种方法得到的识别率相差不多,几乎一样.综合表 2

的数据我们得到,用加权 kNN 方法得到的平均识别率仍为 84.72 %,而本文方法得到的为 92.52 %.这说明若前后测试样本之间存在序列关系,本文提出的序列的加权 kNN 方法,因利用了这种序列关系所给予的信息,识别率更高.

5 结束语

本文提出了一种序列的加权 kNN 分类方法,它采用贝叶斯联合后验概率作为对测试样本的分类依据.实际上,这个贝叶斯联合后验概率,不仅包含有测试样本的 k 个近邻点所提供的类别信息,而且还包含了前后测试样本间内在的流形结构信息.因此,在序列样本分类的场合,与加权 kNN 分类方法相比,该方法具有更加可靠的分类结果.

参考文献:

- [1] T M Cover, P E Hart. Nearest neighbor pattern classification [J]. IEEE Trans. on Information Theory, 1967, 13 (1) : 21 - 27.
- [2] Y Yang, X Lin. A re-examination of text categorization methods [A]. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. New York: ACM, 1999, 42 - 49.
- [3] Li Baoli, Chen Yuzhong, Yu Shiwen. A comparative study on automatic categorization methods for Chinese search engine [A]. Proceedings of the Eighth Joint International Computer Conference [C]. Hangzhou: Zhejiang University Press, 2002. 117 - 120.
- [4] G Gora, A Wojna. A classifier combining rule induction and k-NN method with automated selection of optimal neighbourhood [A]. Proceedings of the Thirteenth European Conference on Machine Learning [C]. Heidelberg: Springer Berlin, 2002, 2430: 111 - 123.
- [5] C D 'Amato, D Malerba, F Esposito, et al. Extending the k-nearest neighbour classification algorithm to symbolic objects [A]. Atti del Convegno Intermedio della Societ  Italiana di Statistica "Analisi Statistica Multivariata per le scienze economiche-sociali, le scienze naturali e la tecnologia" [C]. Italia: Napoli, 2003.
- [6] W Hechenbichler, K Schliep. Weighted k-nearest-neighbor techniques and ordinal classification [OL]. <http://epub.ub.uni-muenchen.de/1769/>, 2007-4-10/2008-9-12.
- [7] 陈振洲,李磊,姚正安.基于 SVM 的特征加权 kNN 算法 [J]. 中山大学学报(自然科学版), 2005, 44 (1) : 17 - 20. Chen Zhen-zhou, Li Lei, Yao Zheng-an. Feature-weighted k-nearest neighbor algorithm with SVM [J]. Acta Scientiarum Naturalium Universitatis Sunyatseni, 2005, 44 (1) : 17 - 20. (in Chinese)

- [8] 刘 明,袁保宗,唐晓芳. 证据理论 k -NN 规则中确定相似度参数的新方法[J]. 电子学报, 2005, 33(4) : 766 - 768
Liu Ming, Yuan Bao-zong, Tang Xiaofang. A new approach to determine the similarity parameters in evidence-theoretic k -NN rule[J]. Acta Electronica Sinica. 2005, 33(4) : 766 - 768. (in Chinese)
- [9] D P Vivencio, E R Hruschka, M C Nicoletti, et al. Feature-weighted k -nearest neighbor classifier[A]. Proceedings of the IEEE Symposium on Foundations of Computational Intelligence [C], Washington DC, USA : IEEE Communications Society, 2007. 481 - 486.
- [10] 孙 岩,吕世聘,唐一源. 无先序条件约束的 KNN 算法[J]. 小型微型计算机系统, 2008, 29(4) : 682 - 686.
Sun Yan, Lv Shi-p in, Tang Yi-yuan. No previous ordering for k NN algorithm [J]. Journal of Chinese Computer Systems, 2008, 29(4) : 682 - 686. (in Chinese)
- [11] C Shan, S Gong, P W McOwan. Dynamic facial expression recognition using a Bayesian temporal manifold model [A]. Proceedings of British Machine Vision Conference [C]. UK: Edinburgh, 2006, 1 : 297 - 306.
- [12] R O Duda, P E Hart, D G Stork 著. 李宏东, 姚天翔, 等译. 模式分类(第二版) [M]. 北京 : 机械工业出版社, 2006. 16 - 21.
- [13] T Kanade, J Cohn, Y Tian. Comprehensive database for facial

expression analysis [A]. Proceeding of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG '00) [C]. Washington, DC, USA : IEEE Computer Society, 2000. 46 - 53.

- [14] J B Tenenbaum, V de Silva, J C Langford. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290 : 2319 - 2323.

作者简介 :



朱明早 男, 讲师. 1974 年生于湖南慈利县, 2005 年毕业于中南大学信息科学与工程学院, 获硕士学位, 其后在湖南文理学院电气与信息工程学院任教. 现为中南大学信息科学与工程学院博士生, 主要从事模式识别、计算机视觉方面的研究.

E-mail : zhumh. 123 @163. com



罗大庸 男, 教授、博士生导师. 1944 年 10 月生于湖南长沙, 主要从事模式识别、计算机视觉、智能控制、信息融合技术等方面的研究工作.

E-mail : dyluo @mail. csu. edu. cn