基于自产生投票的手写汉字识别

邵允学 [1]　王春恒 [1]　肖柏华 [1]　张琳波 [2]

摘　要　在模式识别领域, 投票策略是非常有效的, 而且已被成功应用到人脸检测、识别等领域. 然而, 在手写汉字识别 (Handwritten Chinese character recognition, HCCR) 中, 由于类别集很大、训练样本少等特点, 现有的很多分类器集成方法方法都很难直接应用于此领域. 本文提出一种自产生式投票的方法, 该方法通过事先学习得到的参数集产生一个测试集合, 然后用一个分类器去识别测试集合中的每个样本, 得到属于各个类别的概率, 最后通过加权投票得到识别结果. 实验结果表明, 本文提出的方法是实用和有效的.

关键词　手写汉字识别, 自产生投票, 改进二次判别函数, 线密度归一化

引用格式　邵允学, 王春恒, 肖柏华, 张琳波. 基于自产生投票的手写汉字识别. 自动化学报, 2013, **39**(4): 450−454

DOI　10.3724/SP.J.1004.2013.00450

Handwritten Chinese Character Recognition Based on Self-generation Voting

SHAO Yun-Xue[1]　WANG Chun-Heng[1]　XIAO Bai-Hua[1]
ZHANG Lin-Bo[2]

Abstract　Voting strategy is very useful in pattern recognition and it has been successfully applied in many applications like face detection and recognition. However, these state-of-the-art methods are infeasible or unsuitable for handwritten Chinese character recognition (HCCR) because of the problem′s characteristics. In this paper, a self-generation voting based method is proposed for further improving the recognition rate in handwritten Chinese character recognition. This method learns a set of parameters first for generating a set of samples from the test sample, and then classify these generated samples using a baseline classifier. At last, it gives the recognition result by voting. Experimental results on two databases show that the proposed method is effective and useful in handwritten Chinese character recognition systems.

Key words　Handwritten Chinese character recognition (HCCR), self-generation voting, modified quadratic discriminant function (MQDF), line density equalization

Citation　Yun-Xue Shao, Chun-Heng Wang, Bai-Hua Xiao, Lin-Bo Zhang. Handwritten Chinese character recognition based on self-generation Voting. *Acta Automatica Sinica*, 2013, **39**(4): 450−454

收稿日期 2012-03-09　录用日期 2012-06-15
Manuscript received March 9, 2012; accepted June 15, 2012
国家自然科学基金 (61172103, 60933010, 60835001) 资助
Supported by National Natural Science Foundation of China (61172103, 60933010, 60835001)
本文责任编委 刘成林
Recommended by Associate Editor LIU Cheng-Lin
1. 中国科学院自动化研究所复杂系统管理与控制国家重点实验室 北京 100190
2. 中国交通运输部科学研究院 北京 100029
1. State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190　2. China Academy of Transportation Sciences, Beijing 100029
该文的英文版同时发表在*Acta Automatica Sinica*, vol. 39, no. 4, pp. 450−454, 2013.

离线手写汉字识别问题 (Handwritten Chinese character recognition, HCCR) 已经被许多研究人员进行了大量且深入的研究, 并且取得了很多成果[1−10]. 例如: 1) 在归一化阶段, 基于线密度均衡的非线性归一化方法[1−2] 的提出; 2) 特征提取阶段, 笔画方向特征[3−5] 的提出; 3) 特征降维阶段, 线性判别分析 (Linear discrimination analysis, LDA) 的应用; 4) 分类阶段, 改进二次判别函数 (Modified quadratic discriminant function, MQDF)[3] 的提出. 本文的基准分类器用到了这几个技术. 由于类别集很大、训练样本少等特性, 很多方法像 ANN (Artifieial neural network), SVM (Support vector machine), Bagging 或 Boosting 都很难直接应用到手写汉字识别中. 然而, 这些方法中的一些好的思想对于解决这个问题是非常有启发的. 本文就利用了投票的思想来进一步提升识别性能. 本文提出的自产生式投票方法, 首先根据预先学习得到的一个参数集产生一个虚拟测试集, 然后利用基准分类器识别测试集合中的每个样本, 得到属于各个类别的概率, 最后通过加权投票得到识别结果. 在两个数据集上的实验结果表明, 该方法具有较好的效果. 而且, 由于并行处理技术的快速发展, 该方法具有的并行性将会受到广泛关注.

本文的安排如下: 首先, 第 1 节介绍了自产生投票的框架, 然后, 在第 2 节介绍了自产生方法和参数学习方法. 第 3 节给出了实验结果. 第 4 节是总结.

1　自产生投票

Bagging 或 Boosting 是两种非常有效的提升分类性能的投票方法. Bagging 是 Breiman[11] 提出的, 给定一个大小为 n 的训练集 D, Bagging 方法通过从 D 中有放回的均匀采样得到 m 个新的大小为 $n_p > n$ 的训练样本集 D_i. 在每个训练集 D_i 上可以得到一个分类模型 m_i. 最终通过投票策略给出分类结果. Boosting 则注重分类错误的样本, 将分类错误的训练样本的权重提高, 并依据修改后的样本权重来生成新的训练样本空间, 并用来训练下一个个体分类器. 每个弱分类器的权重是其在训练集上的正确率. 这两个方法的共同点都是通过学习多个弱分类器 $h_t(x)$ 和对应的权重 α (Bagging 的权重都设为 1). 然后通过投票给出最终分类结果.

$$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{1}$$

给定一个待测试的样本, 每个弱分类器给出一个预测结果, 最终的识别结果通过加权投票决定. 本文将这种投票方式称为多分类器投票. 图 1 是多分类器投票策略的示例图.

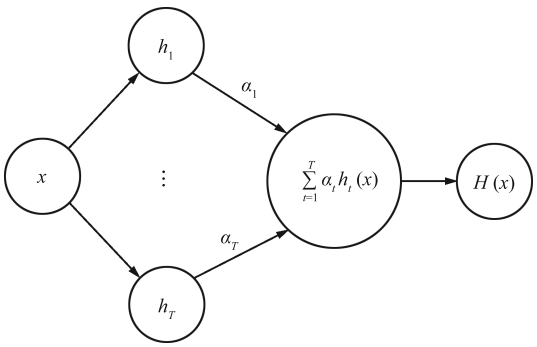

图 1　多分类器投票策略
Fig. 1　The multi-classifier voting strategy

在手写汉字识别中, 由于训练样本少、类别集很大等特点, 类似 Bagging 或 Boosting 的多分类器投票策略很难应用到这个问题中. 怎样才能将投票策略应用到手写汉字识别中? 假设现在只有一个分类器 $h(x)$, 对于一个待测试样本 x, 如果能从这个测试样本产生出一个测试样本集 $\{x_1, \cdots, x_T\}$, 那么可以用分类器 $h(x)$ 来分类每一个产生的样本, 最终的分类结果就可以由投票产生. 本文将这种使用一个分类器, 分类每个由原样本产生的样本, 最终通过这些分类结果投票产生最终结果的策略叫做自产生投票. 自产生投票需要根据测试样本产生一个测试集 $\{x_1, \cdots, x_T\}$ 并且测试集中每个测试样本的投票权重 $\{\alpha_1, \cdots, \alpha_T\}$ 也需要事先学习得到.

$$H(x) = \sum_{t=1}^{T} \alpha_t h(x_t) \tag{2}$$

图 2 是自产生投票策略的示例图.

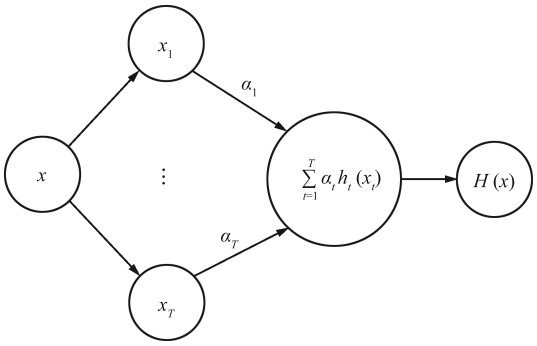

图 2　自产生投票策略

Fig. 2　The self-generation voting strategy

2　基于自产生投票的手写汉字识别

为了使用自产生投票策略. 首先, 需要确定从一个样本产生一个样本集的方法. 产生的样本需要与原样本属于同一类并且不能太相似; 其次, 需要确定一个能够选择小且完备的产生样本子集和对应权重的参数学习方法. 学习得到的子集冗余度小、互补性强, 这样可以使得投票速度更快, 效果更好; 最后, 采用投票策略给出识别结果.

2.1　自产生方法

在手写汉字识别中, 文献 [12] 提出了一种从一个训练样本产生大量虚拟训练样本的自产生模型. 他们的实验结果表明, 产生的虚拟样本能够有效地解决训练样本不足的问题. 本文采用此方法来从测试样本中产生虚拟的测试样本, 这个产生方法能够保证所产生的样本属于同一个类别, 而且可以通过调节参数, 控制产生样本与原始样本的差距.

假设 $f(x, y)$ 是原始字符图像, $g(u, v)$ 是产生的字符图像. 从原始图像 $f(x, y)$ 到产生图像 $g(u, v)$ 的映射函数为

$$\begin{aligned} u &= u(x, y) = w_n(d_1, b_1(x)) + k_1 b_2(y) + c_1 \\ v &= v(x, y) = w_n(d_2, b_2(y)) + k_2 b_1(x) + c_2 \end{aligned} \tag{3}$$

其中, k_1 和 k_2 控制修剪坡度, d_1 和 d_2 控制局部缩放的程度, c_1 和 c_2 是为了将映射的中心调节到图像的中心, b_1 和 b_2 将原始坐标线性映射到区间 $[0, 1]$, w_n 是非线性扭曲函数, 用来对局部图像进行扭曲变形.

实验中, 所有的参数设置都和文献 [12] 相同: k_1 和 k_2 分别从区间 $[-0.17, 0.17]$ 和 $[-0.20, 0.20]$ 中随机选取, $d_1 \neq 0$ 和 $d_2 \neq 0$, 从区间 $[-1.6, 1.6]$ 中随机选取, 使用下面两个扭曲函数 w_1 和 w_2, 使用 w_1 和 w_2 的概率为 0.8 和 0.2, 可以产生更多形状的虚拟样本.

$$w_1(d, t) = \frac{1 - \mathrm{e}^{-dt}}{1 - \mathrm{e}^{-d}} \tag{4}$$

$$w_2(d, t) = \begin{cases} 0.5 w_1(d, 2t), & 0 \leq t \leq 0.5 \\ 0.5 + 0.5 w_1(-d, 2(t - 0.5)), & 0.5 < t \leq 1 \end{cases} \tag{5}$$

图 3 给出了根据不同 d_1 ($d_2 = 0.01, k_1 = 0, k_2 = 0, w_n = w_1$) 所产生样本的示例, 中间矩形框内的字符是原始图像. 从左上到右下, d_1 的值依次是 $-1.6 \sim 1.6$, 间隔 0.1. 可以看出 $|d_1|$ 越大, 字符的形变就越大, 太极端的参数值产生的样本将不利于投票识别, 根据这个准则确定自产生模型中参数范围, 从图 3 还可以看出, 某个参数的微小变化, 对产生样本的形状影响不大.

图 3　根据不同 d_1 产生的样本.

Fig. 3　Generated samples with different d_1

图 4 给出了一些产生样本的例子, 所用参数都是在各个范围内随机选择的. 第一列是原始图像, 其他列是根据不同参数产生的样本. 可以看出产生的样本和原样本属于同一类且形状不同.

图 4　产生样本实例

Fig. 4　Some generated samples

2.2　参数学习方法

根据第 2.1 节描述的产生模型, 一个参数集 $P_i\{d_1, d_2, k_1, k_2, w_n\}$ 对应一个产生样本 $g_i(u, v)$.

$$P_i \rightarrow g_i(u, v) \tag{6}$$

不同的参数集集合所对应的投票结果会有所不同, 本文随机产生 100 个参数集 $S_p\{P_1\{d_1^1, d_2^1, k_1^1, k_2^1, w_n^1\}, \cdots, P_{100}\{d_1^{100}, d_2^{100}, k_1^{100}, k_2^{100}, w_n^{100}\}\}$ 用于训练分类器和参数学习.

参数学习的目标是从这 100 个随机产生的集合中, 学习一个子集 $P\{P_1, \cdots, P_T\}$ 和对应的投票权重, 使得在验证集

上分类结果最好. 算法 1 给出了本文采用的学习算法, 采用贪心法从候选集合 S_p 和 S_α 中依次选取 P_i 和 α_i, S_α 是投票权重的候选集合.

算法 1 中, T 是要选择的参数集的个数, $\emptyset$ 为空集合, $CT_0\{(\varphi_1^0),\cdots,(\varphi_n^0)\}$ 是验证集, φ_i^0 是从样本 f_i 上提取的特征, 局部变量 $best\_p$ 是第 t 次循环中得到的最优参数集, $best\_p=\Phi$ 意味着没有参数集对已学习的集合在验证集上有帮助.

算法 1. 参数集和投票权重学习算法

输入. P,α

初始化. $P=\emptyset,\alpha=\emptyset,best\_acc=0$

for $t=1:T$ **do**

  $best\_p=\emptyset$;

  For $p\in S_p$ **do**

    1) 使用参数集 p 从样本 f_i 产生 g_i;

    2) 从产生的样本 g_i 上提取特征 φ_i^t;

    3) 将特征 φ_i^t 放进集合 CT_{t-1}, 得到集合 CT_t, $CT_t=\{(\varphi_1^0,\cdots,\varphi_1^t),\cdots,(\varphi_n^0,\cdots,\varphi_n^t)\}$;

    for $\alpha\in S_\alpha$ **do**

      1) 令 $\alpha_t=\alpha$, 得到集合 $\alpha\{\alpha_1,\cdots,\alpha_t\}$;

      2) 计算在验证集上当前的识别率 cur_acc, 分类器为 $H(f_j)=h(\varphi_j^0)+\sum_{i=1}^t\alpha_t h(\varphi_j^i)$;

      3) **if** $cur\_acc>best\_acc$ **then**

        $best\_acc=cur\_acc, best\_p=p, best\_\alpha=\alpha$;

      end

    end

  end

  if $best\_p=\emptyset$ **then**

    Break

    else

      1) 使用参数集 $best\_p$ 提取特征 φ_i^t 且放进集合 CT_{t-1}, 得到集合 $CT_t=\{(\varphi_1^0,\cdots,\varphi_1^t),\cdots,(\varphi_n^0,\cdots,\varphi_n^t)\}$;

      2) 令 $\alpha_t=best\_\alpha$ 得到集合 $\alpha\{\alpha_1,\cdots,\alpha_t\}$;

      3) 令 $P_t=best\_p$ 得到集合 $P\{P_1,\cdots P_t\}$;

    end

  end

end

3 实验分析

为了验证本文方法的有效性, 在两个常用的数据集 CASIA 和 CASIA-HWDB1.1[13] 上做了充分的实验. CASIA 数据集由中国科学院自动化研究所收集, 包含常用的 3 755 个汉字, 每个汉字有 300 个样本. CASIA-HWDB1.1 数据集由中国科学院自动化研究所模式识别实验室收集, 包含 3 755 个汉字及 171 个字符, 本文实验只用了 3 755 个汉字. 在这两个数据集上都是将前 250 个作为训练数据, 后 50 个作为测试数据. 图 5 给出了两个数据集上的样本实例, 前三行是 CASIA 中的样本, 后三行是 CASIA-HWDB1.1 数据集上的样本.

在参数学习时, 训练样本又被分为训练和验证两个部分, 每类中前 240 个作为参数学习的训练集, 后10个作为验证集. 权重的候选集为 α $\{0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1\}$, 权重的很小变化对投票结果的影响很小, 许多基于复合距离[14] 的方法都采用类似的权重候选集.

图 5 两个数据集上的样本实例

Fig. 5 Some samples in the two databases

对于每个原始样本和产生的样本, 采用基于线密度均衡化[1] 的归一化方法, 将其归一化到 64×64, 然后在归一化的图像上, 提取 512 维的 8 方向梯度特征[5], 再用 LDA 将每个特征降到 256 维. 实验中用到的分类器是 MQDF2[3], MQDF2 中主要的参数为使用的特征向量个数 N_v. 在算法 1 的 $h(x)$ 分类器中, N_v 设置为 50. 对于每个待测试样本 x, 根据学习得到的参数 $P\{P_1,\cdots,P_T\}$, 产生测试样本集 $\{x_1,\cdots,x_T\}$. 首先用采用粗分类器和 MQDF2 分类原始样本 x, 得到前 10 个候选, 然后根据 x 以及 $\{x_1,\cdots,x_T\}$ 到候选类别的距离加权和最小的类别作为最终的分类结果.

假设产生一个虚拟样本需要时间为 $TIME_n$, 粗分类时间为 $TIME_c$, 最终投票的时间为 $TIME_v$, 则识别一个样本总共的时间为 $(T\times TIME_n+TIME_c+TIME_v)$. 而不采用本文方法需要的时间为 $(TIME_n+TIME_c)$. 在一台普配置为双核、4G 内存的 PC 机上, $TIME_n$ 约为 1.4 ms, $TIME_c$ 约为 4.8 ms, $TIME_v$ 约为 3.7 ms (10 个候选且 $T=20$). 可以看出, 最费时的是产生样本阶段, 虽然该方法非常适合并行处理, 但在以后的工作中, 还需要考虑怎样能够快速地产生一个测试样本集.

首先, 只用原始的训练样本来训练分类器, 评价在测试集上的结果. 表 1 给出了不同 N_v 对应的识别结果, 在两个数据集上, 最好的识别结果都是当 $N_v=50$ 时, 当 N_v 变大时, 识别率降低, 这是由于训练样本不够.

表 1 没有使用产生样本时的识别结果 (%)

Table 1 The recognition rates of the base-line classifier without using generated samples (%)

N_v	CASIA	CASIA-HWDB1.1
30	97.87	87.24
50	**97.97**	**87.41**
70	97.95	87.07
90	97.90	87.07

然后, 采用原始样本和产生的样本来训练分类器, 评价在测试集上的结果. 表 2 给出了不同 N_v 对应的识别结果, 当 N_v 变大时, 识别率增加, 这表明产生的虚拟样本能够有效缓解训练样本不足的问题.

表 2 使用产生样本作为训练样本的识别结果 (%)

Table 2 The recognition rates of the base-line classifier using generated samples for training (%)

N_v	CASIA	CASIA-HWDB1.1
30	97.75	87.04
50	97.98	87.74
70	98.07	**88.13**
90	**98.10**	88.06

最后, 评价本文方法在两个数据集上的性能. 表 3 和表 4 给出了不同 T 和 N_v 对应的识别结果, $T = 0$ 意味着没有产生样本参与投票. 从表 3 和表 4 可以看出, 本文方法比原始不采用投票的方法要好, N_v 增加时, 识别率也增加, 这与表 2 一致.

图 6 是 $N_v = 50$ 时, 不同 T 时的识别率, 左图是 CASIA 上的结果, 右图是 CASIA-HWDB1.1 上的结果. 可以看出, 当 T 增加到 20 左右时, 识别率的增加变得缓慢, 这说明不需要太多的虚拟样本参与投票, 就可以得到较好的结果.

表 3 本文方法在 CASIA 上的结果 (%)

Table 3 The recognition rates of the proposed method on database CASIA (%)

N_v	$T=0$	$T=5$	$T=10$	$T=15$	$T=20$	$T=25$
30	97.75	98.16	98.41	98.56	98.62	98.64
50	97.98	98.28	98.48	98.62	98.67	98.69
70	98.07	98.31	98.49	98.63	98.68	**98.70**
90	98.10	98.27	98.48	98.61	98.65	98.66

表 4 本文方法在 CASIA-HWDB1.1 上的结果 (%)

Table 4 The recognition rates of the proposed method on database CASIA-HWDB1.1 (%)

N_v	$T=0$	$T=5$	$T=10$	$T=15$	$T=20$	$T=25$
30	87.04	88.47	89.68	90.24	90.38	90.50
50	87.74	88.91	89.83	90.36	90.52	**90.65**
70	88.13	88.89	89.83	90.40	90.51	90.59
90	88.06	88.95	89.85	90.39	90.49	90.63

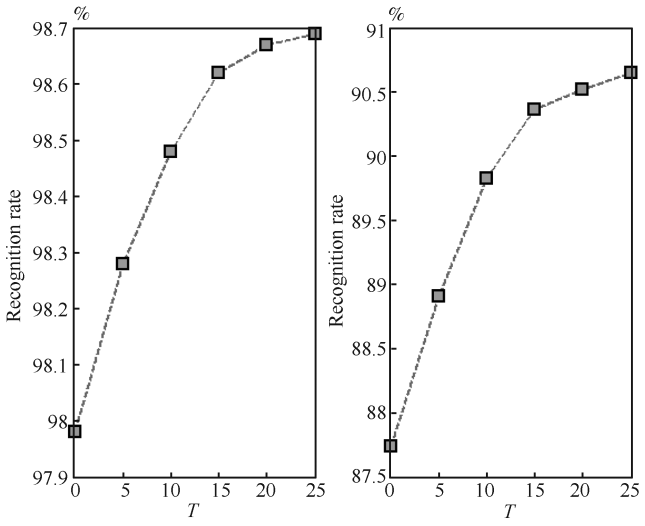

图 6 不同 T ($N_v = 50$) 对应的识别结果

Fig. 6 The recognition rates with varying T ($N_v = 50$) on the two databases

4 结论

本文提出了基于自产生投票的手写汉字识别方法. 该方法通过学习一组产生样本的参数集和对应的投票权重, 对测试样本进行自产生投票. 在两个常用的数据集上的测试结果表明, 该方法是有效的, 且仅需要少于 20 个样本的投票, 分类性能就可以得到很好的提升.

References

1 Yamada H, Yamamoto K, Saito T. A nonlinear normalization method for handprinted Kanji character recognition-line density equalization. *Pattern Recognition*, 1990, **23**(9): 1023−1029

2 Tsukumo J, Tanaka H. Classification of handprinted Chinese characters using nonlinear normalization and correlation methods. In: Proceedings of the 9th International Conference Pattern Recognition. Rome, Italy: IEEE, 1988. 168−171

3 Kimura F, Takashina K, Tsuruoka S, Miyake Y. Modified quadratic discriminant functions and the application to Chinese character recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1987, **9**(1): 149−153

4 Shi M, Fujisawa Y, Wakabayashi T, Kimura F. Handwritten numeral recognition using gradient and curvature of gray scale image. *Pattern Recognition*, 2002, **35**(10): 2051−2059

5 Liu C L, Nakashima K, Sako H, Fujisawa H. Handwritten digit recognition: investigation of normalization and feature extraction techniques. *Pattern Recognition*, 2004, **37**(2): 265−279

6 Gao T F, Liu C L. High accuracy handwritten Chinese character recognition using LDA-based compound distances. *Pattern Recognition*, 2008, **41**(11): 3442−3451

7 Liu C L. High accuracy handwritten Chinese character recognition using quadratic classifiers with discriminative feature extraction. In: Proceedings of the 18th International Conference on Pattern Recognition. Washington DC: IEEE, 2006. 942−945

8 Liu C L, Sako H, Fujisawa H. Handwritten Chinese character recognition: alternatives to nonlinear normalization. In: Proceedings of the 7th International Conference on Document Analysis and Recognition. Edinburgh, UK: IEEE, 2003. 524−528

9 Leung K C, Leung C H. Recognition of handwritten Chinese characters by critical region analysis. *Pattern Recognition*, 2010, **43**(3); 949−961

10 Xu B, Huang K Z, Liu C L. Similar handwritten Chinese characters recognition by critical region selection based on average symmetric uncertainty. In: Proceedings of the 2010 International Conference on Frontiers in Handwriting Recognition. Kolkata, India: IEEE, 2010. 527−532

11 Breiman L. Bagging predictors. *Machine Learning*, 1996, **24**(2): 123−140

12 Leung K C, Leung C H. Recognition of handwritten Chinese characters by combining regularization, Fisher′s discriminant and distorted sample generation. In: Proceedings of the 10th International Conference on Document Analysis and Recognition. Barcelona, Spain: IEEE, 2009. 1026−1030

13 Liu C L, Yin F, Wang D H, Wang Q F. CASIA online and offline Chinese handwriting databases. In: Proceedings of the 2011 International Conference on Document Analysis and Recognition. Beijing, China: IEEE, 2011. 37−41

14 Suzuki M, Ohmachi S, Kato N, Aso H, Nemoto Y. A discrimination method of similar characters using compound Mahalanobis function. *Trans. IEICE Jpn*, 1997, **J80-D-II**(10): 2752−2760

邵允学 中国科学院自动化研究所博士研究生. 主要研究方向为图像处理和模式识别. E-mail: yunxue.shao@ia.ac.cn
(**SHAO Yun-Xue** Ph. D. candidate at the Institute of Automation, Chinese Academy of Sciences. His research interest covers image processing and pattern recognition.)

王春恒 中国科学院自动化研究所研究员. 主要研究方向为图像处理和模式识别. 本文通信作者. E-mail: chunheng.wang@ia.ac.cn
(**WANG Chun-Heng** Professor at the Institute of Automation, Chinese Academy of Sciences. His research interest covers image processing and pattern recognition. Corresponding author of this paper.)

肖柏华 中国科学院自动化研究所研究员. 主要研究方向为图像处理和模式识别. E-mail: baihua.xiao@ia.ac.cn
(**XIAO Bai-Hua** Professor at the Institute of Automation, Chinese Academy of Sciences. His research interest covers image processing and pattern recognition.)

张琳波 中国交通运输部科学研究院助理研究员. 主要研究方向为计算机视觉和模式识别. E-mail: lumbard.zhang@gmail.com
(**ZHANG Lin-Bo** Research assistant at China Academy of Transportation Sciences. His research interest covers computer vision and pattern recognition.)