doi:10.16018/j.cnki.cn32-1650/n.201604009

文字识别中特征与相似度度量的研究

李　杰,方木云
(安徽工业大学 计算机科学与技术学院,安徽 马鞍山　243002)

摘要:在大样本测试集下国内现有成熟的 OCR 识别软件的首位识别准确率为 95% ~97%之间,在准确率和方法上仍有提升和改进的空间。提出一种基于概率特征和结构特征融合的自适应文字识别算法,模拟人类学习的模式,通过对训练样本的不断学习去构建汉字在测量空间的概率分布矩阵,然后比对原始图像和标准汉字库中汉字的概率分布矩阵的相似度来达到汉字分类的效果。其中相似度度量准则是从矩阵空间的结构和概率 2 个角度出发去构建的,充分考虑了结构模式识别和统计模式识别的优缺点。实验结果显示算法在训练样本下的首位识别正确率可以达到 99.66%,在 1 623 张非训练样本文字图像下的首位识别正确率可以达到 99.13%,在 5 515 张非训练样本文字图像下的首位识别正确率可以达到 98.57%。可以证明提出的相似度度量方法在文字识别中的有效性。

关键词:概率特征;结构特征;相似度;文字识别
中图分类号:TP311　　**文献标识码**:A　　**文章编号**:1671-5322(2016)04-0042-05

在计算机视觉和模式识别领域,特征是很重要的研究对象,目标分类、识别、图像检索等问题,都可以归结为对目标进行特征提取,利用不同的相似度定义与先验信息进行比较,得到结果的过程[1]。相似性反映了两个物体或两个特征之间的关系程度,是衡量样本之间相似程度的一个重要指标[2]。根据研究对象的不同,相似度又可以细化如向量相似度、系统相似度、形状相似度等[3-8]。本文通过模拟人类学习过程,对样本进行不断的学习与优化,获取每个汉字在矩阵空间内的分布信息,然后从矩阵空间中节点的权值及结构角度构建相似度函数实现对汉字的分类。

1　训练学习

常见的汉字尺寸为 20×20 ~25×25 像素之间。若把汉字的每个像素点看作一个特征单元,那么字符的像素尺寸越小,组成汉字图像的特征数目就越少,字符原有的有效信息就越有可能损失。若选择的训练样本尺寸偏小,会造成训练得到的节点权值失真,不能很好地区分相似字;若选择的训练样本尺寸偏大,又会造成训练及识别时间的不必要增加。实验比较 32×32 结构与 36×36 结构[9],识别准确率并未太多变化,相反由于计算维度变小,识别的速度变得更快,所以从存储容量、计算速度、识别效率 3 方面考虑,本文采用了 32×32 的矩阵空间对汉字归一化后的信息进行存储。文中随机选取了 3 500 个汉字进行实验,每个汉字 W_i 分别使用 1、120、240 个样本进行训练,得到不同训练样本下汉字 W_i 的节点权重矩阵 $F_i^k(k=1,120,240)$。在对未知汉字样本 M 分类时,会依据权重矩阵 F_i^k 对样本 M 进行相似度计算,并依据其大小作出最终决策。

2　特征选择

由于每个汉字的信息是存储在 32×32 的矩阵空间中的,在对 2 个矩阵进行相似度计算时,需要进行 1 024 维度的向量计算,对于单个的文字识别尚可接受,但是若对每个汉字分类都进行这

收稿日期:2016-05-20
基金项目:安徽省自然科学基金资助项目(1308085QF113)
作者简介:李杰(1990—),男,安徽滁州人,硕士生,主要研究方向为图像处理与数据挖掘。

样的计算,那将耗费大量时间。所以,为了缩短识别时间,本文在相似度计算时采用 2 级分类的方式[10],即对待测样本归一化后的二值矩阵按照整行、整列单元选取并取行列单元的交集节点作为特征单元进行相似度计算,当相似度结果大于一个指定阈值时,保留分类汉字,反之,则舍弃。初级和 2 级分类的区别在于选择相似度计算的特征单元不同,初级分类是从局部特征的角度进行识别[11-13],选取的特征单元数目相对较少,可以筛选掉与待测样本相似度较大的汉字分类;而 2 级分类是从全局特征的角度进行识别,所选的特征单元相对较多,区分相似汉字的能力更强。

2.1　初级分类特征选择

对于归一化为 32 ×32 的待分类样本 M,当特征维数选择 1 024 时,则需进行 1 024 维度的相似度计算。从准确性上说所选维度越多计算效果越佳,但高维度计算带来的是识别的实时性不强。从人类的认知习惯可知,为了权衡识别的时间及准确性,人们对汉字的识别往往会先根据字体的结构比如左右、上下,或是其它特定的偏旁部首进行初步的筛选,然后根据汉字的其它特征与候选字再次进行对比,所以本文初级分类选择了 8 种特征单元进行组合($X \times Y$ 构造分别为 32 ×3、32 ×6、32 ×11、32 ×16、3 ×32、6 ×32、11 ×32、16 ×32)和实验分析。实验中,当测试集为训练样本时,如果 X 轴上选择 0 ~31 单元,Y 轴选择 0 ~2 单元时,初级识别率 top1 可达到 99. 6%,当 X 轴上选择 0 ~31,Y 轴选择 0 ~15 单元时,对应的识别率为 99. 17%;当测试集为非训练样本时,如果 X 轴上选择 0 ~31,Y 轴选择 0 ~2 单元时,初级识别率 top1 可达到 99. 13%, X 轴上选择 0 ~31,Y 轴选择 0 ~15 单元时,对应的识别率为 98. 52%。由此可见,并非选择的特征单元越多就能达到更好的分类效果,由于初级分类只是对所有样本进行初步筛选,所以特征单元不必选择太多。

2.2　2 级分类特征选择

伴随着分类级别的增加,筛选剩下的汉字数量逐步减小,分类结果间的差异性逐渐减小,此时需要适当地增加特征数目来提高分类间的差异性。实验所得,经过初级分类后首位准确率已经达到 95%,且经过初级分类后,满足条件的汉字数目明显减少许多。为了尽量区分汉字间的差异性,文中 2 级分类的特征单元选择了所有矩阵单元节点。

3　相似度计算

相似度测度一般有两种方法:距离测度法和相似性函数法[14]。距离准则的计算量相对于相似度准则要小一些,但是前者在描述两个向量的相似程度时存在精度不足,且对异常情况下的文字图像识别效果不佳。如当一个待测汉字样本的前景特征节点数量较少时,它与非正确汉字的距离也比较小,不能很好地区别正确与非正确汉字;而当一个字的特征较多时,即使它与正确汉字的模板很接近,其计算出的距离仍然比较大。这种对于放大和缩小没有很好的鲁棒性[15]决定了距离度量法判定结果的不稳定性。本文从待测样本与模板样本的矩阵节点概率和节点结构两个角度提出一种新的相似度判别准则来衡量样本与分类之间的相似程度。具体方法描述如下:

对于待测样本 M 与已知汉字 W_i 分类,令:

S:待测样本 M 的前景节点集合

P_i:分类 W_i 的前景节点集合

C_1:S 与 P_i 的交集节点集合

C_2:S 与 P_i 的差集节点集合

T_1:C_1 对应在 F_i^k 上的和

T_2:P_i 对应在 F_i^k 上的和

定义待测样本 M 与已知汉字 W_i 分类的相似度评价函数为:

$$\begin{aligned}\mathrm{sim}_i(M,W_i) = \\ u_1 \times \mathrm{sum}(C_1)/(\mathrm{sum}(C_1) + \\ \mathrm{sum}(C_2)) + u_2 \times T_1/T_2\end{aligned} \tag{1}$$

其中 u_1、u_2 为影响因子,分别表示矩阵间节点概率的相似度和矩阵节点结构相似度的权值,并且 $u_1 + u_2 = 1$。实验中 u_1、u_2 选取了 10 组数据,当 $u_1 = 0.5$、$u_2 = 0.5$ 时其分类效果最佳。所以,取 $u_1 = 0.5$、$u_2 = 0.5$ 进行实验,并根据得到的相似度大小进行排序,即可得到 M 所属的汉字分类。

通过初级、2 级分类后,训练样本集的首位准确率最高可达 99. 58%,非训练样本集的首位准确率最高可达 99. 13%。从准确率来说,非训练样本识别的准确率与训练样本的准确率相当接近,而且随着训练样本的数目增加,识别的准确率处于递增的状态。

4　实验结果分析

本文共进行 3 组对比实验,其中实验 1 是在

训练样本尺寸为 *32* × *32* 像素情况下,通过训练不同数量的样本,探讨不同尺寸测试样本在不同训练样本下的识别准确率;实验 2 主要研究不同的特征组合方式对于文字识别速度的影响;实验 3 选择了 11 张随机截取的文字图片,每张图片上的文字数量不一,分别采用市场上几款较为成熟的软件进行分析比较,以证明文中所提相似度度量方法的有效性。具体结果如下:

实验 1:

该组实验共有 4 个测试集,分别为:

(1)3 500 张训练样本图像,图像尺寸为 32 × 32 像素

(2)1 623 张随机文本图像,图像尺寸为 27 × 27 像素

(3)1 623 张随机文本图像,图像尺寸为 24 × 24 像素

(4)1 623 张随机文本图像,图像尺寸为 21 × 21 像素

归一化后均选择位于 $X[0,2]$,$Y[0,31]$ 范围内的矩阵节点进行相似度计算,每个测试集分别在训练样本为 1、120、240 时进行测试。

表 1 给出了 4 个测试集分别在 3 种不同训练样本数目下的首位识别正确率的对比情况。

图 1 为 4 个测试集对应的首位正确率趋势。

表 1 训练样本及非训练样本在学习样本为 1、120、240 下的首位正确率比较

Table 1 Comparison of the first correct rate between the training samples and the non training samples in the learning samples such as 1,120,240

序号	测试样本	训练样本								
		1			120			240		
		正确	错误	正确率/%	正确	错误	正确率/%	正确	错误	正确率/%
1	3 500	3 075	425	87.86	3 493	7	99.8	3 488	12	99.66
2	1 623	1 315	308	81.02	1 549	74	95.44	1 609	14	99.13
3	1 623	1 295	328	79.79	1 498	125	92.30	1 593	30	98.15
4	1 623	1 198	425	73.81	1 400	223	86.26	1 509	114	92.97

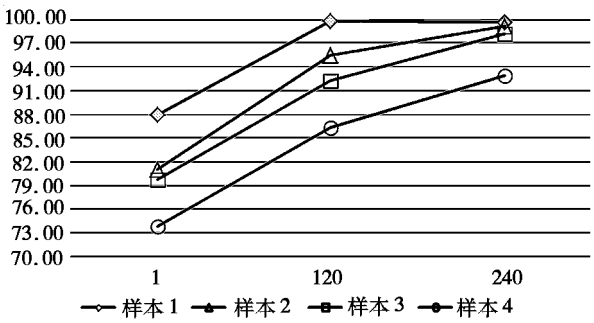

图 1 测试集在不同训练样本下的识别率

Fig. 1 Recognition rate of test set under different training samples

由实验 1 结果可以看出:

(1)对于相同测试集,随着训练样本的增加,识别准确率会有比较明显的提升。

(2)在相同训练样本下,随着测试集图像尺寸越接近训练样本的尺寸,算法识别准确率越高。

(3)训练样本测试集的识别准确率比非训练样本的识别准确率高。

(4)随着训练样本的增加,测试集的整体相似度方差逐渐减小,首位识别正确率逐渐提升。

实验 2:

选取 2 组样本进行测试,测试环境的训练样本均为 240,采用 2 级分类。

第 1 组:从 3 500 个标准汉字的训练样本中每个汉字选取 1 个样本,共计 3 500 张测试图片;

第 2 组:随机获取的 1 623 张尺寸为 27 × 27 的文字图像。

实验结果如表 2、表 3 所示:

第 1 组(训练样本):

表 2 对训练样本在采用不同特征组合下首位准确率比较

Table 2 Comparison of the first accurate rate of the training samples under different combination of features

初级分类		2 级分类		错识数	准确率/%
X 轴特征	*Y* 轴特征	*X* 轴特征	*Y* 轴特征		
(0,31)	(0,2)	(0,31)	(0,31)	14	99.58
(0,31)	(0,5)	(0,31)	(0,31)	16	99.54
(0,31)	(0,10)	(0,31)	(0,31)	20	99.43
(0,31)	(0,15)	(0,31)	(0,31)	28	99.20
(0,2)	(0,31)	(0,31)	(0,31)	40	98.86
(0,5)	(0,31)	(0,31)	(0,31)	19	99.4
(0,10)	(0,31)	(0,31)	(0,31)	38	98.91
(0,15)	(0,31)	(0,31)	(0,31)	42	98.8

第 2 组(非训练样本):

表 3　对非训练样本在采用不同特征组合下首位准确率比较

Table 3　Comparison of the first accurate rate of non training samples under different combination of features

初级分类		2 级分类		错识数	准确率/%
X 轴特征	Y 轴特征	X 轴特征	Y 轴特征		
(0,31)	(0,2)	(0,31)	(0,31)	14	99.13
(0,31)	(0,5)	(0,31)	(0,31)	16	99.01
(0,31)	(0,10)	(0,31)	(0,31)	20	98.77
(0,31)	(0,15)	(0,31)	(0,31)	24	98.52
(0,2)	(0,31)	(0,31)	(0,31)	16	99.01
(0,5)	(0,31)	(0,31)	(0,31)	21	98.71
(0,10)	(0,31)	(0,31)	(0,31)	22	98.64
(0,15)	(0,31)	(0,31)	(0,31)	19	98.83

由实验 2 可以得出:

(1)随着特征单元数目变多,识别准确率并没有相应地提高;相反当初级分类特征单元较少时,识别的准确率更高。

(2)训练样本测试集的识别准确率比非训练样本的识别准确率高。

(3)不同的组合方式得到的识别准确率是有区别的;在本实验中上下组合比左右组合分类效果更好。

实验 3:

随机获取 23 张文档图片,共 5 515 个汉字,分别与现有两款较为成熟的 OCR 软件的首位识别准确率进行对比。实验结果如表 4 所示:

由实验 3 结果可以得出:

文中所提出的相似度度量方法在汉字分类中的效果较同类产品有较为明显的提升。

表 4　不同软件识别测试比较

Table 2　Comparison of different software identification tests

测试图片	软件			
	OCR 1	OCR 2	文中结果	汉字总数
1	45	38	12	608
2	32	30	3	595
3	40	20	3	505
4	42	35	4	649
5	34	31	10	451
6	25	20	10	380
7	17	14	4	273
8	19	15	9	219
9	1	2	7	648
10	3	1	5	358
11	12	10	12	829
Sum	270	216	79	5 515
正确率/%	95.10	96.08	98.57	

5　结束语

实验 1 使用 3 500 张训练样本及 1 623 张未知文字样本构建 4 组数据进行对比,结果表明通过训练的样本越多,文字识别的准确率越高,并且测试的文字图像尺寸越接近训练样本尺寸,文字识别准确率越高;实验 2 结果表明不同特征单元的组合获得的文字识别率不一样。本文算法的准确性是建立在对大量样本的学习之上的,随着学习的文字图像数目的增加,文字识别的准确性是逐步提升的,在图片样本出现黏结、缺失、粗细变化等情况下也具有较好的鲁棒性且误识率较低。这种不断优化的方式也很符合人类的认知习惯,即学习得越多,对某一个汉字分类的结果就越准确。综上所述,本文提出的基于结构和概率特征融合的自适应方法在文字识别中具有较为有效的作用。

参考文献:

[1] 李珊珊. 计算机视觉中特征与相似性度量研究[D]. 合肥:中国科学技术大学,2010.

[2] 史忠植. 高级人工智能[M]. 北京:科学出版社,2011.

[3] 焦利明,于伟,罗均平,等. 向量相似度在雷达目标识别中的应用[J]. 火控雷达技术, 2006,35(2):78-81.

[4] 田晓东,刘忠. 基于形状相似度的水下目标识别算法[J]. 声学技术, 2007,26(3):493-497.

[5] 王婷. 本体相似度研究[J]. 电脑知识与技术,2007,1(6):139-141.

[6] 许瑞丽,徐泽水. 区间数相似度研究[J]. 数学的实践与认识,2007,37(24):1-8.

[7] 戚晓明,陆桂华,吴志勇,等. 水文相似度及其应用[J]. 水利学报,2007,38(3):355-360.

[8] 郑丽萍,李光耀,梁永全,等. 本体中概念相似度的计算[J]. 计算机工程与应用,2006,42(30):25-27.

[9] 任金昌,赵荣椿,张炜. 一种快速有效的印刷体文字识别算法[J]. 中国图象图形学报:A 辑,2001,6(10):1 011-1 015.

[10] 卢婷. 基于 AdaBoost 的分类器学习算法比较研究[D]. 上海:华东理工大学,2014.
[11] 李庆佑,刘宏申,李兴. 联机手写连笔字符识别特征提取算法优化[J]. 人类工效学,2014,20(5):56-59.
[12] 胡龙灿,孙怀义,樊爱军. 手写字符识别中的特征提取研究[J]. 自动化与仪器仪表,2013(2):15-17.
[13] LING H, SOATTO S, 2007. Proximity Distribution Kernels for Geometric Context in Category Recognition[C]//IEEE International Conference on Computer Vision. IEEE, 2007:1-8.
[14] 郭亮勇,王国海. 联机手写字符识别技术研究与实现[J]. 软件导刊,2013,12(5):145-146.
[15] 冯伟兴,贺波,王臣业,等. Visual C + + 数字图像模式识别技术详解[M]. 2 版. 北京:机械工业出版社,2013.

Research on Feature and Similarity Measurement in Character Recognition

LI Jie, FANG Muyun
(School of Computer Science and Technology, Anhui University of Technology, Maanshan Anhui 243002, China)

Abstract: In the large sample test set, the first recognition accuracy of the existing mature OCR recognition software is 95% ~ 97%. There is still a space for improvement and improvement in accuracy and method. An adaptive character recognition algorithm based on the fusion of probability feature and structure feature is proposed. By simulating the model of human learning, we construct the probability distribution matrix of Chinese characters in the measurement space through continuous learning of training samples, and then compare the similarity between the original image and the probability distribution matrix of Chinese characters in the standard Chinese character library to achieve the effect of Chinese character classification. The similarity measurement criterion is constructed from two angles of the structure and probability of matrix space, and the advantages and disadvantages of structural pattern recognition and statistical pattern recognition are fully considered. The experimental results show that the algorithm can achieve the first recognition accuracy rate of 99.66% in the training samples. The first recognition accuracy of the 1 623 non-training sample text images can reach 99.13%. The first recognition accuracy of the 5 515 non-training sample text images can reach 98.57%. It can be proved that the proposed similarity measure method is effective in word recognition.

Keywords: probability feature; structure feature; similarity; character recognition

(责任编辑:李华云)