

视频检索在汉字识别中的应用研究

桂丹萍, 陈佳祥, 何红生
(集美大学, 福建 厦门 361021)

摘要:传统的OCR技术在汉字识别领域趋于成熟,对背景清晰的正体汉字有很高的识别正确率,然而当汉字图片在复杂背景中或经旋转、加噪处理后,OCR软件的识别正确率大大下降。当今有关视频检索的研究正在快速发展中,其中一种行之有效的方法是通过提取模板视频的关键帧及其特征向量,应用聚类算法形成关键字,并通过快速的检索算法来实现匹配。创新性地将该模型应用到汉字识别研究中,通过大量实验数据的研究发现,该模型在上述情况中相对于传统的OCR技术优势明显,在未来实际应用中具有广阔的前景。

关键词:汉字识别;视频检索模型;SIFT特征;KMEANS;TFIDF

中图分类号:TP391.43

文献标识码:A

文章编号:1673-629X(2010)10-0207-04

Application Research of Video Retrieval Model on Chinese Character Recognition

GUI Dan-ping, CHEN Jia-xiang, HE Hong-sheng
(Jimei University, Xiamen 361021, China)

Abstract: Traditional OCR has achieved a degree of maturity in the field of Chinese character recognition, which obtains a high recognition accuracy on Chinese character with a clean background and no rotation. However, when images are preprocessed in a complex background with low quality like affine transform and addition of noise, its recognition accuracy declined significantly. The current research on video retrieval is growing rapidly, where an effective method is to extract key frames from the video template and their feature vectors, apply clustering algorithm to form keywords, and retrieve the target video through a fast search algorithm. Innovatively apply the model to the study of Chinese character recognition. Through a large number of experimental data, this model outperforms traditional OCR under such variances. Therefore, this model enjoys a good prospect of application in the future.

Key words: chinese character recognition; video retrieval model; SIFT feature; KMEANS; TFIDF

0 引言

随着计算机技术的不断进步,我国的汉字识别研究工作取得了令人瞩目的成果,对实际的国民经济发挥着不可或缺的积极效用。由于汉字固有的复杂结构,数量巨大的字库,使之无论在理论还是在实践上都比少量的西方字符研究困难得多,同时预示着有关西方字符的匹配研究并不适用于汉字识别。1966年IBM公司的Casey和Nagy发表了第一篇关于汉字识别的文章^[1],通过采用模板匹配法识别了1000个印刷体汉字,为随后的国内相关研究做好了铺垫。回顾汉字识别的历程,基于视觉形象思维心理学的分析和基于原始图像的统计模式识别方法是取得对超大字符集

汉字识别的成果的基础^[2]。OCR(Optical Character Recognition)利用各种模式识别算法分析文字形态特征^[3],通过光学输入方式获取文字图像信息,并将其转换为计算机能够统一识别和存储的内码,从而识别判断出汉字的标准编码。

现有的汉字识别OCR软件主要针对扫描的高分辨率图像和不包含字幕制作特效的手写或印刷体文字,识别率可以达到99%以上,然而从图像中提取汉字通常都需要首先定位包含汉字的图像区域,当汉字在字体、大小、对齐方式和排列上变化多端,背景复杂且图像分辨率低时,OCR软件从图像中有效地提取出汉字变得非常困难^[4]。特别是当汉字经旋转处理时,在实验中OCR几乎不能识别出任何汉字。

近些年来,基于内容的视频信号与图像库检索的相关技术研究^[5]得到了该领域研究学者的青睐,在文献^[6]中,作者首先将文本检索的思想应用于视频检索中,通过对视频关键特征的提取达到视频检索的目的。

收稿日期:2010-01-16;修回日期:2010-04-29

基金项目:福建省自然科学基金(2007J0202)

作者简介:桂丹萍(1983-),女,硕士研究生,研究方向为非线性方程、图像处理。

文中借鉴该思想,并对提取的 SIFT 特征进行 KMEANS 聚类,通过对不同 K 值下的情况进行分析,实验结果表明该模型对旋转的汉字识别相较于 OCR 软件优势明显。

1 汉字识别流程及其实现方案

图 1 是文中提出的汉字识别匹配模型,该模型以

当前的视频检索模型为参考,从离线处理和在线检索两种方式进行规划。首先在离线状态生成每个汉字的标准高分辨率图片,同时对每个汉字进行预处理(如噪声、颜色、大小、仿射变换等),生成复杂背景下的测试图片;然后对每一张模板图片运用 SIFT 提取算法提取各个局部特征,通过 KMEANS 聚类生成 K 个关键字,该步骤可以大大减少实验数据量,从而提高识别效率;最后进行 TFIDF 操作生成高维向量索引,为实验提供测试环境。在线的汉字识别流程与离线基本相似,检索过程中将测试图片的权重(TFIDF)与数据库中所有图片的倒序索引向量进行匹配(归一化标量积),并选择相似度最高的模板图片作为测试图片的最佳匹配。

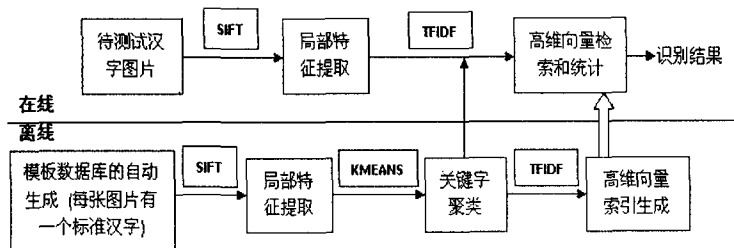


图 1 汉字识别流程图

接下来,以 3 个层次为序分别详细描述模型各个算法的具体实现方案。

1.1 SIFT 特征描述子提取

SIFT 算法由 Lowe 提出^[7],其特征描述子是基于生物视觉理论模型提出的一种新的图像局部特征描述子。SIFT 特征描述子利用特征点领域图像窗口内梯度的方向统计直方图来构造特征描述向量,本实验同样采用高斯差分尺度空间(difference of Gaussian scale space),DoG 算子的计算公式如下:

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) \otimes I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \quad (1)$$

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad (2)$$

式中尺度空间 $L(x, y, \sigma)$ 是尺度可变高斯函数与二维图片 $I(x, y)$ 的卷积, σ 是尺度坐标, k 是分离相邻尺度的乘法因子。

SIFT 特征描述子的生成过程如图 2 所示,对每一个采样点和它同尺度的 8 个相邻点和上下相邻点对应的 9×2 个点共 26 个点进行比较(见图 2a),检测并获取 DoG 尺度空间局部极值点。然后构建尺度空间所需的参数,本实验采用改进的 SIFT 特征提取算法,尺度空间坐标、octave 坐标和 sub-level 坐标参数设置如下:

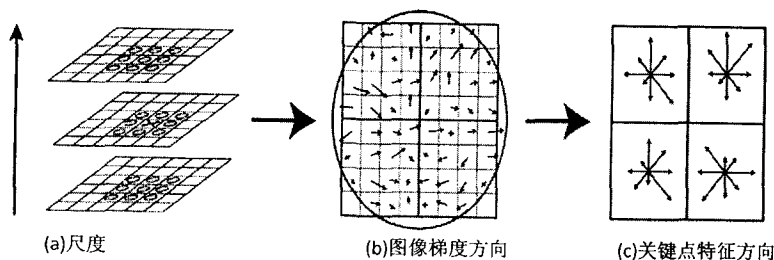


图 2 SIFT 描述子生成示意图

$$\sigma_n = 0.5, \sigma_0 = 1.6, o_{\min} = -1, \text{octaveNum} = -$$

1(程序自动计算 octave 的层数), $S = 3$

$$\text{其中 } \sigma \text{ 和 } o, S \text{ 的关系为: } \sigma(o, s) = \sigma_0 2^{o+s/S},$$

$$o \in o_{\min} + [0, \dots, O-1], s \in [0, \dots, S-1]$$

实验表明改进的算法提取出的向量个数比传统的高出 50%。以关键点为中心取 8×8 的窗口,计算窗口内每个像素点的梯度和方向(见图 2b),在每 4×4 的小块上计算 8 个方向的梯度方向直方图,绘制每个梯度方向的累加值,即可形成一个种子点。一个关键点由 2×2 共 4 个种子点组成(见图 2c),每个种子点有 8 个方向向量信息。 (x, y) 处的梯度模值和方向计算公式如下:

$$\begin{aligned} m(x, y) &= \{(L(x+1, y) - L(x-1, y))^2 + \\ & (L(x, y+1) - L(x, y-1))^2\}^{\frac{1}{2}} \end{aligned} \quad (3)$$

$$\theta(x, y) = \arctan\left(\frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)}\right) \quad (4)$$

本实验中算法采用 16×16 大小窗口,即每个特征点用一个 $4 \times 4 \times 8$ 维的特征向量来描述。当然也可以使用经过改进和扩展的 SIFT 算法,这类算法的代表有 PCA-SIFT 和 GLOH^[8,9],鉴于传统 SIFT 特征(128 维)对图像仿射变换、复杂背景的高鲁棒性,采用传统 SIFT 算法可以很好地达到本实验的目标。

1.2 KMEANS 聚类

通过对上述生成的 SIFT 描述子进行 KMEANS 算法聚类^[10]得到关键字词汇表^[11],KMEANS 算法以 k 为参数,将 n 个 SIFT 描述子划分为 k 个聚类,其中同一聚类中的描述子相似度较高;而不同聚类中的描述子相似度较小。聚类相似度利用各聚类中对象的均值所获得一个“中心对象”(引力中心)获得。

KMEANS 算法的操作流程如下:首先从 n 个数据对象中任意选 k 个对象作为初始聚类中心,而对剩余的其它对象,根据其与其每个聚类中心的相似度(距离),分别将它们赋给其最相似的(聚类中心所代表的)聚类;然后重新计算每个聚类中所有对象的均值,不断重复这一过程直到标准测度函数开始收敛为止。一般都采用均方差作为标准测度函数。KMEANS 聚类对本实验的汉字识别作用明显,它最终将生成少数的关键字($k \ll n$),在 k 值不影响识别正确率的前提下可以大大提高汉字识别的时间效率。

1.3 TFIDF 算法

TFIDF (term frequency/inverse document frequency) 算法^[12]被公认为信息检索中最重要的发明,它描述了单个“单词”(即关键字)与特定“文档”(本实验中即为汉字图片)的相关性,它是一个在特定条件下的关键词概率分布的交叉熵(Kullback - Leibler Divergence),假定每张汉字图片由一个 k 维向量($V_d = (t_1, \dots, t_i, \dots, t_k)^T$)表示,其计算公式如下:

$$t_i = \frac{n_{id}}{n_d} * \log \frac{N}{n_i} \quad (5)$$

其中 n_{id} 是单词 i 在文档 d 中的出现次数, n_d 是文档 d 中所含单词的总数, n_i 是单词 i 在整个数据库中的总次数, N 是整个数据库中的文档总数。TFIDF 权重信息由两部分组成:单词出现频率 $TF (n_{id}/n_d)$ 和文档出现频率的倒数 $IDF (\log(N/n_i))$, 最后的汉字识别通过角度余弦值来衡量每张汉字图片的相似性。

2 实验及其结果分析

为了验证该模型在汉字识别方面的性能,文中随机选取了 500 个汉字进行测试匹配,对模板图片预处理操作,如仿射变换、噪声(加性高斯白噪声)加入、复杂背景等,每个汉字测试图片由 4 个经不同预处理后的图片组成(如图 3 所示),模板数据库中共有 3 种字体(长城宋体、黑体和楷体-GB2312),生成的 SIFT 描述子总数为 178867,测试图片共 6000($3 \times 500 \times 4$)张,在该模型下测试图片的识别匹配结果如表 1 所示。

表 1 汉字识别模型结果

字体 \ k	识别正确率				
	10000	15000	20000	25000	30000
长城宋体	52.80%	55.25%	61.05%	62.75%	56.45%
黑体	59.65%	60.45%	68.25%	69.15%	64.00%
楷体-GB2312	49.25%	52.35%	58.45%	59.35%	53.90%

实验对照组选用两款流行的汉字识别 OCR 软件(清华 TH-OCR 9.0 专业版和汉王文本王 文豪 7600),它们对清晰的正体汉字都有非常高的识别正确

率。为便于测试,从相同的测试图片数据库中随机选取了 120 张图片在 OCR 软件中进行识别,其实验结果如表 2 所示。

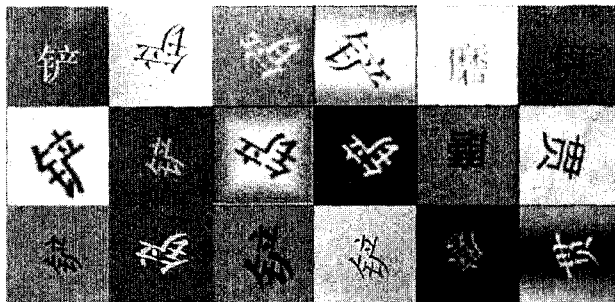


图 3 测试图片示意图(自上而下分别为长城宋体、黑体和楷体-GB2312)

表 2 OCR 软件测试结果

实验组	识别正确率
TH-OCR 9.0	0.83%
文豪 7600	4.2%

从上述实验结果中可以看出,两款 OCR 软件几乎不能识别复杂背景下的旋转汉字,而文中的汉字识别模型在这方面具有较强的鲁棒性,汉字的识别正确率随着 k 的增大(10000 ~ 25000 范围内)而上升(见表 1),当 k 值为 30000 时,识别正确率下降,因为随着 k 的增大,更多的关键字被聚类,此时模糊的 SIFT 描述子聚类到多个新的错误的关键字中,进而某些汉字图片新生成的关键字在识别时覆盖正确的原有关键字的作用,使汉字识别匹配错误。因此当 k 值为模板数据库图片 SIFT 描述子总数的 10% ~ 15% 时,汉字识别的效率最高。

为了进一步验证该模型在图片模糊度较低情况下的识别效率并得出 k 的经验值,做了另一组实验,同样随机选取 500 个黑体汉字进行测试,但对这些图片进行的预处理操作较前一组实验复杂度低,即测试图片更为清晰(如图 4 所示),与实际应用更接近,模板图片生成的 SIFT 描述子总数为 63944 个,不同的 k 值与汉

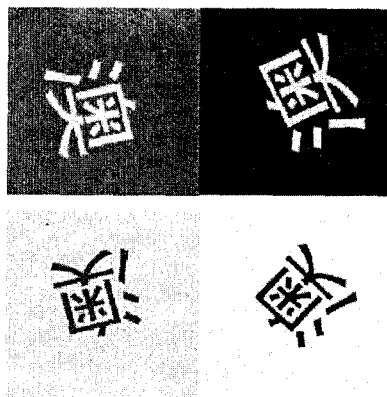


图 4 测试图片示意图

字识别准确率之间的关系如图 5 所示,当 k 取值 6000 (约为描述子总数的 10%) 左右时,汉字的正确识别即接近 90%,因此该模型在汉字识别领域具有广阔的应用前景。

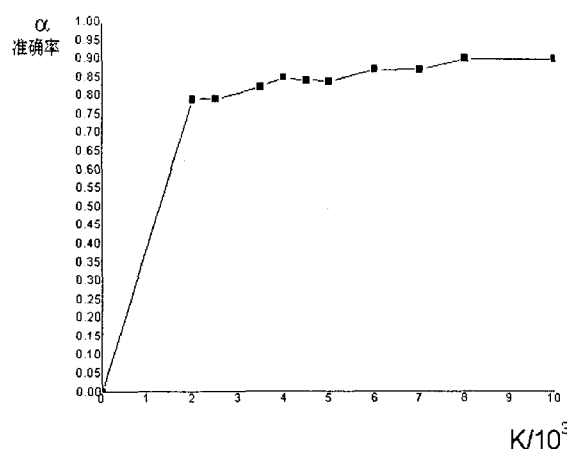


图 5 不同的 k 值与汉字识别正确率之间的关系示意图

3 结束语

文中借鉴了视频检索模型思想,并引入 KMEANS 聚类操作,有效地解决了复杂背景下的旋转汉字难以识别的问题。文中的汉字识别模型对噪声、复杂背景和旋转等字体的识别具有较强的鲁棒性。实验结果表明,在实际应用中该模型的综合性能优于传统的 OCR 汉字识别软件,其研究和应用价值不可忽视。

参考文献:

- [1] Casey R, Nagy G. Recognition of Printed Chinese Characters [J]. IEEE Trans. Electronic Computers, 1996, 15(1): 91 -

101.

- [2] 丁晓青. 汉字识别研究的回顾[J]. 电子学报, 2002(9): 64 - 68.
- [3] Mori S, Suen C Y, Yamamoto K. Historical review of OCR research and development[J]. Proceedings of IEEE, 1992, 80 (7): 1029 - 1058.
- [4] 王 勇, 郑 辉, 胡德文. 图像和视频中的文字获取技术[J]. 中国图象图形学报, 2004, 9(5): 532 - 538.
- [5] 卢汉洁, 孔维新, 廖 明, 等. 基于内容的视频信号与图像库检索中的图像技术[J]. 自动化学报, 2001, 21(1): 56 - 69.
- [6] Sivic J, Zisserman A. Video Google: a text retrieval approach to object matching in videos[C]// Proceedings of International Conference on Computer Vision. Washington, DC: [s. n.], 2003: 1470 - 1477.
- [7] Lowe D G. Distinctive image features from scale - invariant keypoints [J]. International Journal of Computer Vision, 2004, 60 (2): 91 - 110.
- [8] Ke Y, Sukthankar R. PCA - SIFT: A more distinctive representation for local image descriptors[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR' 2004). Washington, DC: [s. n.], 2004: 506 - 513.
- [9] Mikolajczyk K, Schmid C. A performance evaluation of local descriptors[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(10): 1615 - 1630.
- [10] 孙吉贵, 刘 杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48 - 61.
- [11] 彭敦陆, 周傲英. 基于方法聚类的 Web 服务检索技术[J]. 计算机应用, 2007, 27(10): 2365 - 2368.
- [12] 施聪莺, 徐朝军, 杨晓江. TFIDF 算法研究综述[J]. 计算机应用, 2009, 29(6): 167 - 170.

(上接第 206 页)

参考文献:

- [1] SRU: Search and Retrieve via URL (standards, Library of Congress)[EB/OL]. 2007 - 08 - 23. <http://www.loc.gov/standards/sru/index.html>.
- [2] SRW/U (OCLC - Software) [EB/OL]. 2007 - 08 - 23. <http://www.oclc.org/asiapacific/zcn/research/software/srw/default.htm>.
- [3] Eric L M. An introduction to the search/retrieve URL service (SRU)[EB/OL]. 2007 - 08 - 23. <http://www.ariadne.ac.uk/issue40/morgan/>.
- [4] 李春旺, 王小梅, 王 昉, 等. 基于 SRU 的集成服务平台设计与实现[J]. 现代图书情报技术, 2007, 23(10): 12 - 15.
- [5] 王 沛, 冯曼菲. 征服 AJAX - Web2.0 开发技术详解[M]. 北京: 人民邮电出版社, 2006.
- [6] 徐 驰. Ajax 模式在异步交互 Web 环境中的应用[J]. 计算机技术与发展, 2006, 16(11): 228 - 230.
- [7] 杨晓俊. Web2.0 下的 Ajax 及其应用[J]. 计算机与数字工程, 2007, 35(8): 157 - 160.
- [8] 柯昌正, 黄厚宽. Ajax 技术的原理与应用[J]. 铁路计算机应用, 2007(1): 47 - 49.
- [9] 王 东, 孙 彬. 基于 Ajax 的 MVC 框架的改造分析[J]. 计算机应用, 2007(s1): 301 - 303.
- [10] 冉春玉, 童 莹. Ajax 技术及其 Web 开发[J]. 福建电脑, 2007(3): 100 - 101.
- [11] 左伟明. 即用即查 - XML 数据标记语言参考手册[M]. 北京: 人民邮电出版社, 2007.
- [12] 侯要红, 栗松涛. Java XML 应用程序设计[M]. 北京: 机械工业出版社, 2007.
- [13] 蔡 剑, 景 楠. Java 网络程序设计: J2EE (含 1.4 最新功能)[M]. 北京: 清华大学出版社, 2003.