

基于 Gabor-统计特征与 SVM 的文档图像文本检测方法

刘权, 苏海, 苗敏婧

(武汉大学, 武汉 430079)

摘要: **目的** 对文档图像中的文本进行精确检测, 深入研究统计特征对于文字纹理特征分类的影响。**方法** 首先结合 Gabor-统计特征获得文档图像的特征图像, 再应用 SCA 算法提取文本样本和非文本样本, 最后采用 SVM 实现文本检测, 而统计特征的选择使用 Fisher 准则实现。**结果** 依据 Fisher 准则, 逆差距特征对于 Gabor 特征分类的类间离散度最大, 效果最佳。**结论** 针对不同类型的文档图像, 使用 Gabor-逆差距特征能够获得较好的检测效果。

关键词: 文档图像; Gabor-统计特征; SVM; 文本检测

中图分类号: TP391.1 **文献标识码:** A **文章编号:** 1001-3563(2014)23-0100-04

A Text Detection Approach Based on Gabor-statistical Feature and SVM

LIU Quan, SU Hai, MIAO Min-jing

(Wuhan University, Wuhan 430079, China)

ABSTRACT: **Objective** To further improve the quality of document image text detection, in-depth research was performed to analyze how statistical features influenced the classification of text texture. **Methods** First, the document images' feature images were obtained through Gabor-statistical feature, and then the SCA algorithm was applied to extract the text and non-text samples. Finally, SVM was employed to fulfill the text detection. To choose the statistical feature, Fisher criteria were used. **Results** The experiments implied that homogeneity returned the maximum class separation distance according to Fisher criteria and gave the best detection result. **Conclusion** A relatively good detection result could be obtained using Gabor-homogeneity feature when dealing with different types of document images.

KEY WORDS: document images; Gabor-statistical feature; SVM; text detection

文档图像的文本检测即是将图像分为文本区域和非文本区域。图文分离之后能够对文档图像进行深入处理, 例如, 对不同的图像区域采用不同的压缩处理, 或者利用检测结果建立数据库中基于内容的文档图像检索。OCR 技术主要针对文字处理, 对于结构复杂的图像难以达到满意的效果, 因此, 对文本检测方法的研究具有重要意义^[1]。

文档图像文本检测技术基本可以分为自底向上^[2]、自顶向下^[3]和基于纹理的方法等 3 类。自底向上的方法从图像中最小的像素开始, 通过递归将单个像素合

并为区域或者连通成分, 再合并为更大的结构, 自底向上的方法具有较大的灵活性, 但运算量较大; 自顶向下的方法包括 RLS(Run Length Smoothing)^[4]和投影分析(projection profiles)^[5]等方法, 可以处理结构较为单一的科技论文图像, 但是对于结构复杂的图像, 效果并不能令人满意; 基于纹理的方法包括 Gabor 滤波、马尔科夫随机场^[6]等, 其中, Gabor 滤波器是一组有限带通的滤波器组, 具有在空间域和频率域同时取得局部最优化的特性, 而且 Gabor 滤波器的频率和方向表示接近人类视觉系统对于频率和方向的表示, 因此在

收稿日期: 2014-06-06

基金项目: 国家科技支撑计划资助项目(2013BAH03B01); 广东省教育部产学研结合重大专项(2012A090300017)

作者简介: 刘权(1989—), 男, 河北衡水人, 武汉大学硕士研究生, 主攻图像处理和模式识别。

特征提取领域应用非常广泛^[7-10]。文献[11]提出的傅里叶-统计特征,用于视频图像中的文本检测取得了很好的效果,视频图像含有少量人工文本和场景文本,而文档图像中主要为文本,结合文档图像和视频图像的差异,将文字视为一种特殊的纹理,提出Gabor-统计特征与SVM的文档图像的文本检测方法。Gabor滤波采用6个方向、3个尺度的Gabor滤波器,对滤波特征图像采用SCA聚类算法取得文本样本,对不同Gabor-统计特征的文本样本采用Fisher准则评价优劣,并最终应用SVM对Gabor-统计特征图像进行文本检测。

1 基于Gabor与SVM的文档图像文本检测方法

算法详细流程见图1。

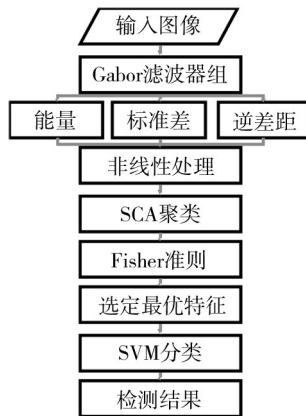


图1 算法流程

Fig.1 Workflow of the algorithm

1.1 Gabor小波

对于给定频率和方向角的Gabor小波为:

$$g(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right) + 2\pi jWx\right] \quad (1)$$

式中, x, y 为像素点坐标; W 为空间频率; σ_x 和 σ_y 为高斯函数的标准差。

对于1幅输入图像 $I(x, y)$,其滤波图像为:

$$GC_{mn}(x, y) = |I(x, y) \cdot g(x, y)| \quad (2)$$

令 U_l 和 U_h 分别表示Gabor小波的最大和最小频率。令 K 表示方向数, S 表示多分辨率分解的尺度数,这里Gabor参数设置采用经验值: $K=6, S=3$,

$$U_l = \sqrt{2}/8, U_h = \sqrt{2}/2。$$

1.2 Gabor-统计特征

统计特征包括能量、能量矩、惯力矩、均值、标准差、三阶中心距和逆差距^[11],这里将文字视为特殊的纹理,因而选取与纹理特征相关性较大的能量、标准差和逆差距作为Gabor-统计特征。其中,能量 E 见式(3),反映图像灰度分布均匀程度和纹理粗细度;标准差 S_{sd} 见式(4),反映纹理变化的波动性;逆差距 H_m ^[12]见式(5),反映图像纹理的均匀性。

$$E = \sum_{x,y} GC_{mn}^2(p, q) \quad (3)$$

$$S_{sd} = \sqrt{\frac{1}{M-1} \sum_{-w/2}^{w/2} \sum_{-w/2}^{w/2} |GC_{mn}(p+i, q+j) - \mu|^2} \quad (4)$$

$$H_m = \sum_{x,y} \frac{1}{1 + (p-q)^2} GC_{mn}^2(p, q) \quad (5)$$

其中, p, q 为像素坐标; i, j 为窗口的坐标; w 为窗口大小; μ 是窗口中像素的均值。

构建特征图像时,对每个滤波后的图像使用窗口大小为 $M=w \times w$ 像素的滑动窗求取每个像素的特征,再采用 $\phi(x)$ 进行非线性变换,见式(6)。

$$\phi(x) = \tanh(\alpha x) = \frac{1 - e^{-2\alpha x}}{1 + e^{-2\alpha x}} \quad (6)$$

式中, α 取经验值0.25。对线性变换后的图像进行高斯低通滤波平滑,实验中高斯窗大小为17像素 $\times$ 17像素。

1.3 SCA聚类和SVM

将SCA聚类应用于文档图像中的文本样本和非文本样本的分类。SCA聚类是一种基于图论的聚类方法,将带权无向图分为2个或者2个以上的最优子图,使子图内部尽量相近,而子图间距离尽量较远,以达到聚类的目的,与K-means等其他聚类方法不同,SCA聚类可以收敛到全局最优解,克服了K-means可能收敛于局部最优的缺陷。文献[13]中假设输入数据点为 $x_1, x_2, \dots, x_n$;聚类数为 k 。SCA聚类步骤如下所述。

- 1) 根据数据点构造相似度矩阵 $S \in R^{n \times n}$ 。
- 2) 计算拉普拉斯矩阵 L 。
- 3) 计算 L 最小的 k 个特征值及其特征向量。
- 4) 将 k 个特征向量组成 $N \times k$ 的矩阵 V 。

5) 通过4)计算 V 的归一化矩阵 U ;

6) 对矩阵 U 中的行向量使用 K-means 算法聚为类。行向量的类别即为 n 个数据点所对应的类别。

对一幅特征图像直接应用 SCA 算法, 需要一个很大的邻接矩阵存储数据, 使得运算量过大。为了降低运算复杂度, 先对特征图像进行采样, 采用 SCA 聚类获得 SVM 的训练样本, 再使用 SVM 分类提高分类效率。

采样过程为: 将图像分为 $50 \text{ 像素} \times 50 \text{ 像素}$ 的子块, 每个子块中取 1 个像素点的特征向量进行 SCA 聚类。每个样本点分别为文本样本或者非文本样本。

实验中采用 RBF 为训练函数进行 SVM 训练, 首先将 SCA 聚类后样本点所属类别与该点特征向量组合为新的特征向量, 并将该向量作为 SVM 的训练样本进行训练。训练之后, 对 Gabor-统计特征的所有特征向量通过训练得到的 SVM 进行分类, 得到最后的分类结果。

1.4 特征选择

Fisher 判别式和 Fisher 准则用来评价不同 Gabor-统计特征, 生成特征向量的性能好坏^[14-15], 针对 1.2 提出的能量、均方差和逆差距等 3 个图像特征的 SCA 文本样本, 应用 Fisher 准则来判定这 3 个特征的优劣, 并进行实验验证。

Fisher 准则要使类间离散度越大越好, 类内离散度越小越好, 其函数定义为 $J_F(w) = (m_1 - m_2)^2 \cdot (S_1^2 + S_2^2)^{-1}$ 。其中, m_1, m_2 为类内均值; S_1, S_2 为类内离散度。

针对不同 Gabor-统计特征图像 SCA 的分类结果, 分别在文本样本和非文本样本中随机选取 1000 个特征向量, 再根据 Fisher 准则计算这 2 类的类内离散度均值 S_1, S_2 和它们的类间离散度均值 S_w , 最终选取 S_w 最大的特征作为提取特征图像的特征。

2 实验与分析

试验采用的文档图像取自 PRImA 开发的文档图像数据库^[16], 这个数据库共有 55 幅文档图像, 包括不同类型的文档图像、结构简单和复杂的图像, 以及不同字体、不同缩进或者其他不同格式的图像。

2.1 Fisher 准则实验

不同 Gabor-统计特征的检测结果见图 2。对图 2

Gabor-统计特征的 SCA 聚类结果应用 Fisher 准则, 结果见表 1。



图 2 不同 Gabor-统计特征的检测结果
Fig.2 Results of different Gabor-statistical features

表 1 Fisher 准则实验
Tab.1 Fisher criteria experiment

特征/Fisher	S_1	S_2	S_w
Energy	2	0.7	3
Std2	10	1	11
Homogeneity	21	9	31

依据 Fisher 准则, 2 类之间的离散度越大, 分类效果越理想, 从表 1 中可以看出, 逆差距特征检测效果最佳, 随后依次是标准差特征和能量特征, 而最终使用 SVM 后的检测结果见图 2。从图 2 b, c, d 中可以看出, 能量特征将全部的图像检测为文本, 均方差特征也将大部分的图像检测为文本, 而逆差距特征则将较少的图像检测为文本。

根据 Fisher 准则的类间离散度最大原则以及最终的实验结果可知, 使用逆差距特征结合 Gabor 特征, 相对于能量特征和均方差特征可以取得较优的检测效果, 因此, 最终选取 Gabor-逆差距特征用于文本检测。

2.2 实验结果

为了进一步检验该算法, 在 PRImA 的文档数据库中选择了一些典型的图像进行测试, 结果见图 3。图 3a 中含有 2 幅不同位置的图像, 1 幅自然场景图像和 1

幅几何图像,而图3c中含有较大范围的图像与较小范围的文字部分。另外,2幅图片中均包含不同的字体与字号。

应用提出算法的实验结果分别见图3b和图3d,可以看出,检测结果都比较理想,文本与非文本的分离比较彻底,但图像部分有一些区域被判定为文字,其原因是图像中总会有一些区域的特征与文本部分的特征相似。



图3 不同图像的检测结果

Fig.3 Results of different images

3 结语

提出的基于Gabor小波和SVM的文本检测算法,能够对文档图像中的文档定位取得较好的检测结果。利用Gabor-统计特征结合Fisher准则,可以看出,逆差距特征能够很好地用于Gabor特征图像的分类,使用SCA算法和SVM相结合的方法提高了特征向量分类的效率,并且对于不同类型的文档图像也可以取得较好的检测效果。下一步可以对检测结果进行形态学处理和连通分析,从而获得更好的文本检测结果。

参考文献:

[1] 黄为,王云宽,李兵,等. 结合笔画方向信息与SVM的英文文字检测[J]. 中国图象图形学报,2011,16(6):989—994.

HUANG Wei, WANG Yun-kuan, LI Bing, et al. English Text Detection Combining Stroke Direction Information with SVM[J]. Journal of Image and Graphics, 2011, 16(6): 989—994.

[2] 王丹,王希常,杨侠. 一种基于梯度差的文档图像文本行检测算法[J]. 微型机与应用,2011,30(18):32—37.

WANG Dan, WANG Xi-chang, YANG Xia. Text Line Detection of Document Images Based on Gradient Difference[J]. Microcomputer & Its Applications, 2011, 30(18): 32—37.

[3] 何立强,刘浩,陈永. 边缘与灰度检测相结合的场景图像文本定位[J]. 中国图象图形学报,2013,18(12):1601—1608.

HE Li-qiang, LIU Hao, CHEN Yong. Text Extraction from Scene Image Based on Edge and Gray-scale Detection Collaboration[J]. Journal of Image and Graphics, 2013, 18(12): 1601—1608.

[4] PRITI REGE P, CHANCHAL A. Text-Image Separation in Document Images Using Boundary/Perimeter Detection[J]. ACEEE, Signal & Image Processing, 2012, 3(1): 10—14.

[5] FU Xiao-ling, LI Xiao-feng. A Chinese Document Layout Analysis Based on Non-text Images[J]. IEEE Computer Society, 2009(1): 326—328.

[6] KUMAR S, GUPTA R, KHANNA N, et al. Text Extraction and Document Image Segmentation Using Matched Wavelets and MRF Model[J]. IEEE Transactions on Image Processing, 2007, 16(8): 2117—2128.

[7] SABARI R S, PEETA B P, RAMAKRISHNAN A G. Gabor Filter Based Block Energy Analysis for Text Extraction from Digital Document Images[J]. IEEE, Computer Society, 2004(4): 233—243.

[8] QIAO Yu-long, LI Meng, LU Zhe-ming, et al. Gabor Filter Based Text Extraction From Digital Document Images[J]. IEEE Computer Society, 2006(6): 297—300.

[9] 韩润萍,孙苏榕,姜玲. 基于Gabor滤波器组的织物瑕疵点检测方法[J]. 计算机工程与应用,2007,43(7):211—214.

HAN Run-ping, SUN Su-rong, JIANG Ling. Scheme for Fabric Defect Detection Based on Gabor Filters[J]. Computer Engineering and Applications, 2007, 43(7): 211—214.

[10] QIAO Yu-long, LU Zhe-ming, SONG Chun-yan, et al. Document Image Segmentation Using Gabor Wavelet and Kernel-based Methods[J]. IEEE, Systems and Control in Aerospace and Astronautics, 2006(5): 451—455.

[11] SHIVAKUMARA P, PHAN T Q, TAN C L. New Fourier-Statistical Features in RGB Space for Video Text Detection[J]. IEEE Transactions on Circuits and System for Video Technology, 2010, 20(11): 1520—1532.

(下转第114页)

- Lett A, 2010, 374(8):1005—1008.
- [9] CAND S E, ROMBERG J, TAO T. Robust Uncertainty Principles: Exact Signal Reconstruction From Highly Incomplete Frequency Information[J]. IEEE Trans Information Theory, 2006, 52(2):489—509.
- [10] HAUPT J, NOWAK R. Signal Reconstruction from Noisy Random Projections[J]. IEEE Trans Information Theory, 2006, 52(9):4036—4048.
- [11] KATKOVNIK V, ASTOLA J. Computational Ghost Imaging: Advanced Compressive Sensing (CS) Technique[J]. SPIE, 2012, 8413:84130N.
- [12] WELSH S S, EDGAR M P, JONATHAN P, et al. Multi-wave-length Compressive Computational Ghost Imaging[J]. SPIE, 2013, 8618:86180I.
- [13] CHEN Jie, GONG Wen-lin, HAN Shen-sheng. Sub-rayleigh Ghost Imaging via Sparsity Constraints Based on a Digital Micro-mirror Device[J]. Phys Lett A, 2013, 377(31):1844—1847.
- [14] KATKOVNIK V, ASTOLA J. Compressive Sensing Computational Ghost Imaging[J]. J Opt Soc Am A, 2012, 29(8):1556—1567.
- [15] WANG Hui, HAN Shen-sheng, KOLOBOV M I. Quantum Limits of Super-resolution of Optical Sparse Objects via Sparsity Constraint[J]. Opt Express, 2012, 20(21):23235—23252.

(上接第93页)

- Electronic Technology, 2012, 35(8):192—194.
- [11] 王素青, 姜维福. 基于 Matlab/Simulink 的 PID 参数整定[J]. 工业控制与应用, 2009, 28(3):24—26.
- WANG Su-qing, JIANG Wei-fu. PID Tuning Based on Matlab/Simulink[J]. Industry Control and Application, 2009, 28(3):24—26.
- [12] 杨梅, 续明进. 基于 MATLAB 的卷筒纸印刷机张力控制系统的建模与仿真[J]. 包装工程, 2011, 32(7):22—25.
- YANG Mei, XU Ming-jin. Modeling and Simulation of Tension Roll Paper Printing Machine Control System Based on MATLAB[J]. Packaging Engineering, 2011, 32(7):22—25.
- [13] 薛定宇. 控制系统计算机辅助设计—MATLAB 语言及应用[M]. 北京:清华大学出版社, 2012.
- XUE Ding-yu. Control System Computer Aided Design—MATLAB Language and Application[M]. Beijing: Tsinghua University Press, 2012.
- [14] 董霞, 陈康宁, 李天石. 机械控制理论基础[M]. 西安:西安交通大学出版社, 2005.
- DONG Xia, CHEN Kang-ning, LI Tian-shi. Mechanical Control Theory[M]. Xi'an: Xi'an Jiaotong University Press, 2005.

(上接第103页)

- [12] 黄剑华, 唐降龙, 刘家锋, 等. 一种基于 Homogeneity 的文本检测新方法[J]. 智能系统学报, 2007, 2(1):69—73.
- HUANG Jian-hua, TANG Xiao-long, LIU Jia-feng, et al. A New Method for Text Detection Based on Homogeneity[J]. CAAI Transactions on Intelligent Systems, 2007, 2(1):69—73.
- [13] CHEN Wen-yen, SONG Yang-qiu, BAI Hong-jie. Parallen Spectral Clustering in Distributed Systems[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 33(3):568—586.
- [14] GRIGORESCU S E, PETKOV N, KRUIZINGA P. Comparison of Texture Features Based on Gabor Filters[J]. IEEE Transactions on Image Processing, 2002, 11(10):1160—1167.
- [15] WANG Lei, JI Hong-bing, SHI Ya. Face Recognition Using Maximum Local Fisher Discriminant Analysis[C]// IEEE International Conference on Image Processing, 2011:1737—1740.
- [16] ANTONACOPOULO S, BRIDSON D, PAPADOPOULOS C, et al. A Realistic Dataset for Performance Evaluation of Document Layout Analysis[C]// IEEE Computer Society, 2009:296—300.