Software Engineering and Applications 软件工程与应用, 2016, 5(6), 303-310
Published Online December 2016 in Hans. http://www.hanspub.org/journal/sea
http://dx.doi.org/10.12677/sea.2016.56035

Hans汉斯

Application of Tensor Voting in Optical Pre-Processing

Xiaofang Shao, Xiaojun Chu

Qingdao Branch of Naval Aeronautical Engineering Institute, Qingdao Shandong
Email: xiaoxiao_0731@163.com

Received: Nov. 22nd, 2016; accepted: Dec. 10th, 2016; published: Dec. 13th, 2016

Abstract

Optical characteristic extraction and de-skew are two difficulties in optical pre-processing. This paper presents a summary of related works and introduces how to apply the tensor voting method in optical pre-processing. The algorithm's flowchart and typical experimental result are demonstrated to show the completion characteristics of the algorithm.

Keywords

Optical Characteristic Extraction, De-Skew, Tensor Voting

张量投票在文字预处理中的应用

邵晓芳，初晓军

海军航空工程学院青岛校区，山东 青岛
Email: xiaoxiao_0731@163.com

收稿日期：2016年11月22日；录用日期：2016年12月10日；发布日期：2016年12月13日

摘　要

在文字预处理过程中，文字特征提取和倾斜校正是两大难点。在分类总结相关工作的基础上，本文介绍

文章引用: 邵晓芳, 初晓军. 张量投票在文字预处理中的应用[J]. 软件工程与应用, 2016, 5(6): 303-310.
http://dx.doi.org/10.12677/sea.2016.56035

了张量投票方法在文字特征提取和倾斜校正中的应用，并展示了该方法的实验处理效果。

关键词

文字特征提取，倾斜校正，张量投票

1. 引言

随着计算机和网络技术的发展，对文字图像进行计算机处理的需求越来越迫切，文字识别(optical character recognition)应运而生，成为办公室自动化、新闻出版、机器翻译、文本挖掘、字音转换[1]等领域中最为理想的输入方法。另外，文字识别后将庞大的黑白点阵图像压缩成机器内部编码，压缩量在100倍以上，对提高通讯容量及速度也是大有好处的。可以说，文字识别是在强大的社会需求推动下发展起来的一种以功能实现为目标的图像处理技术，其基本原理是将输入文字与各个标准文字进行模式匹配，计算类似度(或距离)，将具有最大类似度的标准文字作为识别结果。信息处理领域中使用文字识别技术可以大大提高计算机的使用效率。但是为了进行模式匹配，必须首先对输入的文本图像进行预处理，实现文字检测和文字分割等处理步骤。显然，文字预处理的效果越好，后续的识别越容易。于是，文字预处理随着文字识别技术的发展成为一个比较活跃的研究领域[2]。

2. 文字识别

目前，文字识别技术从识别文字的难度划分，主要分为手写体文字识别和印刷体文字识别；而从识别的文字类型来划分，可以分为汉字、英文、数字三种。但无论是哪种文字识别，其基本处理流程都可以用图1概括。

图1中，输入一般是由扫描仪、数码相机或者其他数字图像获取设备，把打印或写在纸上的文字转换成具有一定灰度值或彩色颜色值的数字采样信号送人计算机得到的原始文字图像。

预处理环节一般包括去噪、倾斜校正(De-skew)、斑点去除(Despeckle)、二值化、平滑、文字检测(版面分析、线表移除)、文字分割、归一化、线段抽取等，但是根据具体处理方法和目的，预处理的各个步骤不仅仅顺序可以互换，有些步骤还可能被替换、合并或省略。预处理方面涉及到的技术或方法有线形归一、非线性归一、细化、骨架化、整形变换和模糊变换等。

特征提取是在预处理的基础上，提取笔画、笔顺、角点等特征，有些方法将一些笔画或字符组合成部件作为新的特征进行检测。

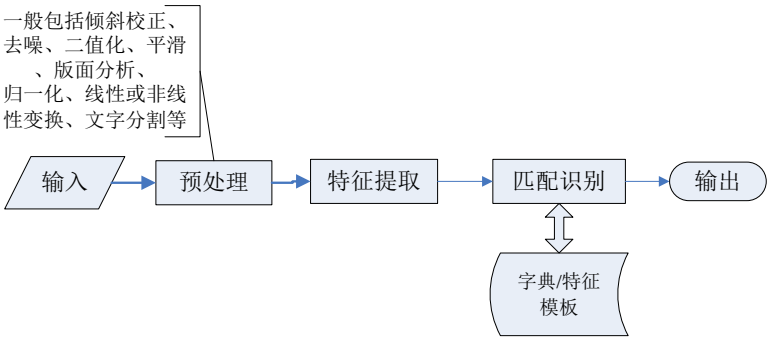


Figure 1. The flowchart for optical character recognition
图 1. 文字识别的一般处理流程

匹配识别过程是采用模式识别和人工智能的理论和方法依据特征提取结果与字典中的字或者特征模板进行模式匹配，匹配结果作为识别结果输出。常用的方法有：关系结构匹配、松弛匹配、逻辑匹配、Bayes 规则、统计偏差、Maharanobis 理论等[2] [3]。

从图 1 中可以看出，对输入图像进行预处理和特征提取是匹配识别的前提，因为文本质量、字体变化、书写风格及工具的变化以及纸张的污损等对于识别的正确率影响都很大。此外，笔者认为，文字识别过程属于模式识别和人工智能的范畴，尽管它也涉及到计算机数字图像处理、模糊数学、组合数学、信息论、自然语言理解、语言文字学、心理学、生物学等相关知识，但严格的数字图像处理环节还是体现在预处理和特征提取这两个步骤上。

虽然文字预处理中的每一部分都有大量的研究成果，但鉴于张量投票算法的应用主要集中在倾斜校正和特征提取这两点上，这里仅对这两点的相关研究成果作以小结。

3. 相关工作

3.1. 特征提取

特征提取通常被认为是文字识别过程中最重要的一环。所提取特征的稳定性和有效性，直接影响识别的性能。特征提取方法虽然形形色色，但根据特征类型可划分为结构特征、统计特征和混合特征三种[4]。

结构特征主要包括字符的笔划端点、交叉点、环、笔划走向、孤立点、笔划关系等特征，常用的方法有骨架扫描法、笔划代码、方向特征法、部件识别法[5]、Gabor 特征法等。骨架扫描法是在细化后的文字骨架上进行，通过顺序扫描计算所有像素的交叉点类型及数目、提取笔划端点等；笔划代码先标定字符像素点的方向代码，再抽取子笔划并生成子笔划点阵，分横、竖、撇、捺四个；方向特征法提取部分字符笔划的相对位置关系特征，可以在笔划边缘点上根据边缘方向在横、竖、撇、捺四个方向属性量化编码或对字符图像顺序扫描生成特征网格，构成一个待识别字符完整的特征矢量；部件识别法把一些文字笔划组合成部件进行整体识别，这种方法实际上是将底层特征组合成了中层特征——部件[5]，然后在部件的基础上完成高层的识别工作；Gabor 特征法需重采样字符点阵并设计 Gabor 滤波器在每个采样点上计算若干 Gabor 特征[6]构成特征矢量。

统计特征是对整个图像或部分图像进行数值测量的计算，梯度特征、笔划密度特征、投影特征、弹性网格特征、几何矩特征等均可划分为统计特征。梯度特征是将目标像素的邻域分成若干扇形区域，对邻域内每一目标像素计算 $\mathrm{d}y/\mathrm{d}x$ 值，再统计每一扇形区域内 $\mathrm{d}y/\mathrm{d}x \neq 0$ 的像素数目排列起来得梯度特征的特征矢量；笔划密度特征从分析字符本身的拓扑结构入手，从不同方向扫描归一化后的字符点阵图像，把各方向扫描线横切字符笔划的次数做叠加即得特征矢量；投影特征分别对图像点阵区域进行 X 轴、Y 轴方向上的投影得到字符像素的统计直方图，字符水平和垂直密度的直方图特征可以较好地反映字符的结构和笔划特征，对整行文字进行投影还可定位文本行的基线；弹性网格特征抽取法是将字符进行横、竖、撇、捺四个方向分解，然后根据笔划分布构造一组非均匀的弹性网格，具体做法是将弹性网格分别作用于待识别字符的四个方向分量上，将字符像素点在网格中的概率分布统计作为字符的特征[4]；几何矩特征利用矩不变量作为特征，对各种干扰适应性强，在线性变换下保持不变[7]。

混合特征将结构特征与统计特征相结合，既吸收了统计特征的优点，又利用了字符的结构信息。代表性的方法有笔划分布的概率密度函数、特征点的高斯密度函数等[4]。

3.2. 倾斜校正

倾斜文档图像是指文献在电子化过程(扫描等)中，由于人为等外界因素影响造成扫描的文档与图像正边成一定角度，即倾斜现象，这种图像称之为倾斜文档图像。在文本扫描过程中，很多原因都可能造成

图像倾斜，给阅读和识别带来很大困难。实践经验表明，3˚以上的倾斜会引起字符的明显变化，大部分识别处理方法难以适应，因而图像倾斜会给文本分割和识别造成很大困难；此外，在表格处理中，图像的倾斜会引起表格识别以及其信息处理困难。倾斜文档图像校正(即倾斜校正)，是指针对倾斜现象，通过各种图像处理技术，校正文档图像中倾斜区域的技术。这是一项重要的预处理技术，其应用非常广泛，尤其在数字化、自动化领域。比如，提高 OCR 识别率从而提高文档自动化处理效率，车牌号码自动识别与交通监视，手写体自动识别，名片自动归类等。倾斜校正可通过某种人机交互手段人工完成，也可由计算机自动完成。

自动倾斜校正可分为整体倾斜校正和局部倾斜校正：整体倾斜校正可以采用统计图像左右两边的平均像素高度，通过计算整体倾斜度来进行校正，这种方法对于像素较多的图像的处理效果明显，而且实现简单快速，但是对于那些已经处理过的单一数字图像并不适用，因为此时的图像一般较小，且笔划较细，由于信息太少统计后的结果并不正确；局部倾斜校正，是认为文档图像呈现的是非一致性倾斜，局部的倾斜特征不一样，针对局部倾斜特点做出的校正，局部倾斜校正也被称为扭曲校正。

倾斜校正的关键是测得输入图形页的倾斜角度，常用方法有外接矩形法、连通体检测法、投影轮廓分析法和基于 Hough 变换的方法[4] [8]。外接矩形法是通过求解文字图像的最小外接矩形(即刚好把所有文字包围在内的矩形框)边的倾角来作为校正依据；连通体检测法中，连通体是一个灰度值相同的像素的集合，这个集合中任意两个像素之间都是 8-近邻关系，如果已知文字行的方向(水平或垂直)，就可以将连通体合并成文字行，并用直线逼近，该直线的倾斜角即为文字行的倾角，对整幅图像的文字行作同样分析，选出出现频率最高的角度作为整幅图像的倾角；投影轮廓分析法是将图像沿其文字行的方向(如水平方向)作投影，并在候选倾斜角度范围内转动图像，直至出现明显的波峰和波谷为止，此时得到的角度即文档的倾斜角；基于 Hough 变换的方法一般流程为先对文档进行连通区域搜索，取各连通区的中点，然后对中点进行 Hough 变换提取出倾角，最后根据 Hough 变换结果计算的倾角进行反向旋转。

倾斜校正的思路还可扩展到解决扭曲校正问题，同样是在图像二值化的基础上，先定位文本行(先垂直扫描图像，找出任何长度大于 T 的垂直线(T 为一个去噪阈值，目的是去除一些不必要的噪声点)，取垂线中点，设为 1，其余作为背景点)，再用最小二乘法拟合曲线，最后选出两条标准曲线进行校正。

4. 基于张量投票的文字预处理

4.1. 特征提取

汉字的复杂性使得细化、骨架化的工作非常艰巨。汉字数量庞大，古今汉字总数约有六万个。常用的《新华字典》收字约 8500 个。另外，中国汉字主要有四种字体，不同字体不但笔划粗细有明显差异，整字的形态和笔划走向也有所变化。不同字体的同一单字，除了拓扑结构基本相同外，其字形、偏旁部首与主体部分的比例、位置，以及笔划的形态、长短、粗细、位置等都有一定差别。总体说来，不同字体的同一个字，其点阵图形是不一样的，而用计算机自动识别时，往往不能把它们看作相同的字[9]。

张量投票方法首先对输入的汉字进行笔划的边界提取，然后通过曲线和交汇点提取形成汉字的骨架和交汇点，从而有利于文字识别。详细处理过程参见文献[10]。简化的流程图如图 2 所示。这里输入是附带极性信息的张量数据，需在投票方程的右端乘上极性信息，如令 q 点的坐标为 (i,j)，则该点的投票结果是通过张量求和得到的，即

$$Vote(i,j)=\sum_{m,n} sign\left(u(i,j)\cdot\hat{e}_1(m,n)\right)\cdot Vote\left(T(m,n),p(m,n,i,j)\right)\Big|_{(m,n)\in N(i,j)} \tag{1}$$

式中，$Vote(i,j)$ 代表投票结果，$N(i,j)$ 代表 (i,j) 的有效邻域，$Vote\left(T(m,n),p(m,n,i,j)\right)$ 代表邻域点

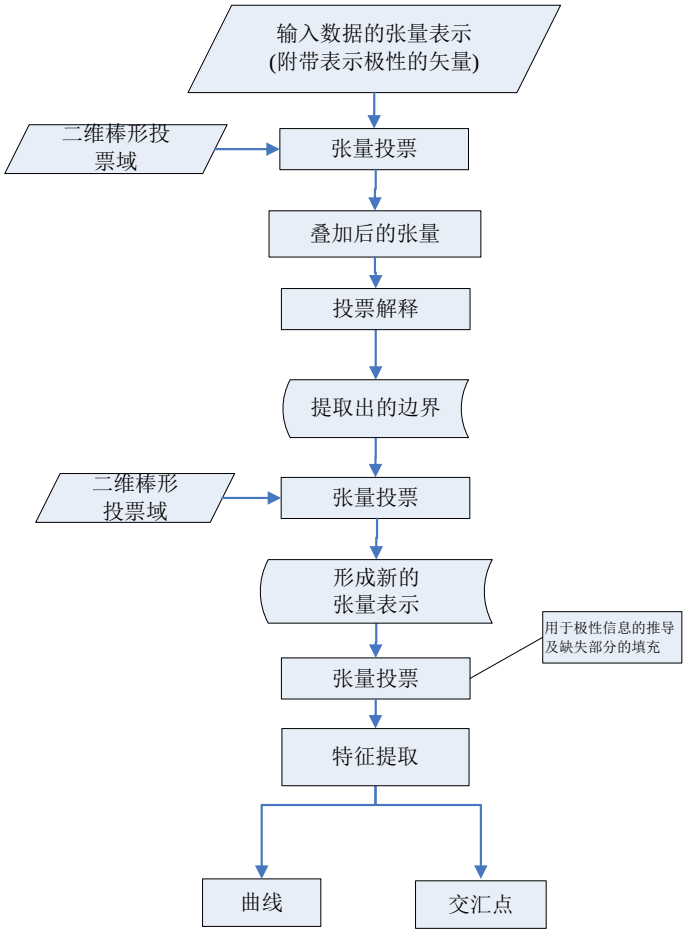


Figure 2. The flowchart for optical character extraction
图 2. 文字特征提取的计算流程

(m,n) 处的张量相对于 (i,j) 的投票值，$u(i,j)$ 表示被投票点正极性方向的单位矢量，而 $\hat{e}_1(m,n)$ 表示投票点的主特征矢量。

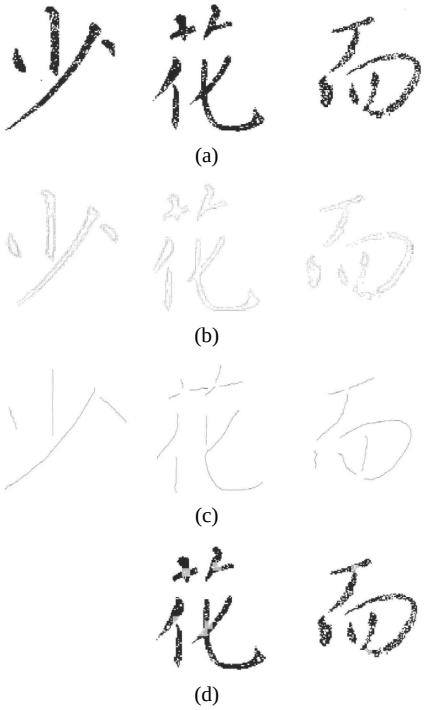
后续的处理过程先是与边界检测过程类似，提取出文字笔划的边界；接下来会应用到线/角点检测方法检测出文字中包含的线段和笔划的交汇点。

张量投票进行文字特征提取的应用示例如图 3 所示。图 3 中，输入的文字为“少”、“花”、“而”三个字，如图 3(a)所示；张量投票算法首先依照前面进行边界提取的计算流程提取出文字的边界，如图 3(b)所示；进一步提取出的骨架如图 3(c)所示，而应用张量投票检测交汇点的结果如图 3(d)所示，其中交汇点用灰色标识。可以看到，边界、骨架和交汇点的信息与人眼的感知结果还是一致的。

4.2. 倾斜校正

张量投票方法利用倾斜文本中的文本行形成曲线的特点，将该问题转化为曲线检测问题：首先提取文本行中各字符的中心点(在去掉过大或过小的笔划之后)，然后通过提取这些中心点形成的曲线计算文本行的倾斜度(倾斜度可以依据 Hough 变换计算)，最后根据计算出的倾角对整篇文字进行反向旋转，从而实现文本校正。这一应用的处理流程如图 4 所示，而实验结果示例如图 5 所示。

图 5 中，图(a)为输入的倾斜图像；图(b)为提取的文字中心点；图(c)为根据文字中心点提取出的文本线；图(d)为校正后的文本。



(a)

(b)

(c)

(d)

Figure 3. The example for Chinese character feature recognition: (a) input character, (b) extracted character contour, (c) skeleton of the character, (d) key points

图 3. 中文文字特征提取示例：(a) 输入文字，(b) 提取出的文字笔划边界，(c) 提取出的文字骨架曲线，(d) 关键点检测

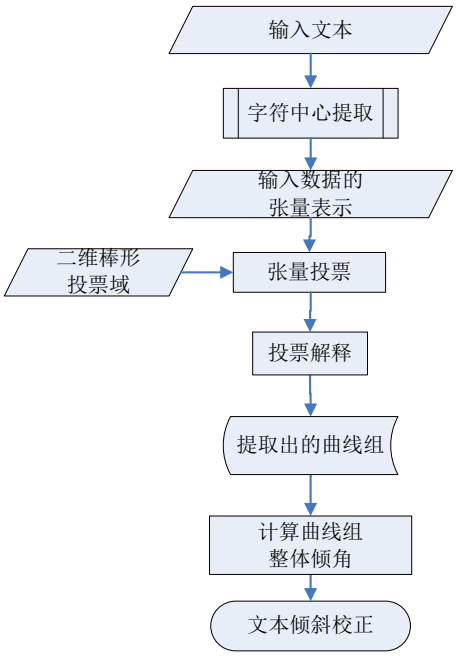
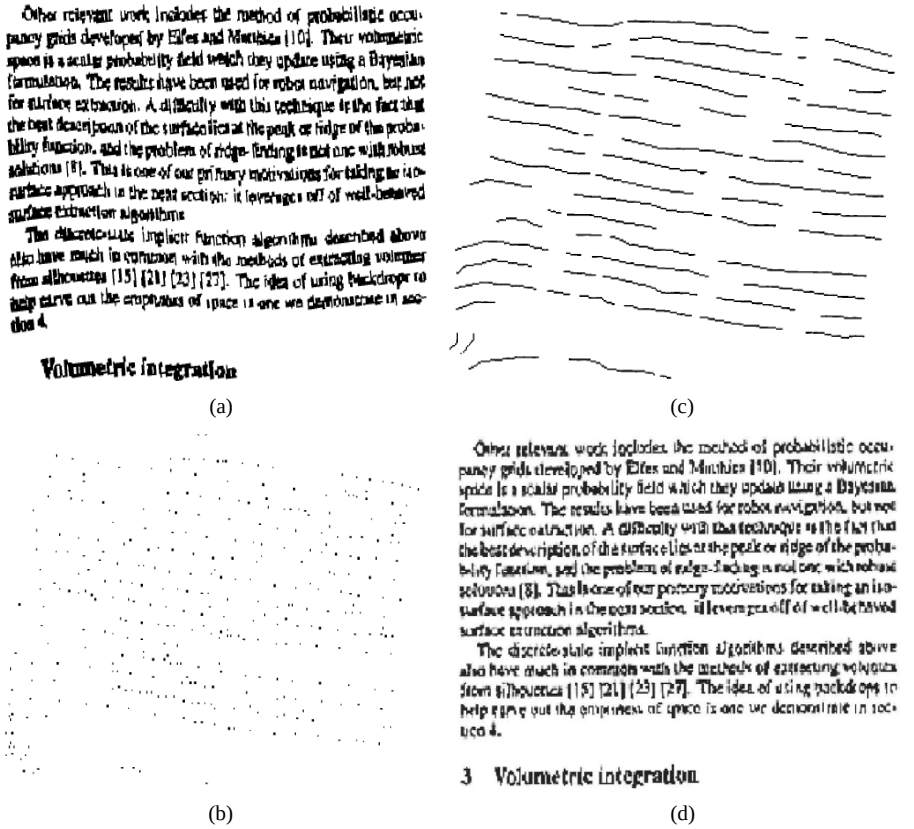


Figure 4. The flowchart for de-skew

图 4. 倾斜校正的计算流程



(a) (c)

(b) (d)

Figure 5. The example for de-skew: (a) input character, (b) extracted character centers, (c) connection line for the characters, (d) regulated characters

图 5. 倾斜文本校正示例：(a) 输入文本，(b) 提取出的文字中心点，(c) 提取出的文本行连线，(d) 校正后的文本

5. 结束语

从本文的描述中可以看到：在文字预处理方面，张量投票方法既可以将汉字笔划包含的特征提取出来，也可以进行文字的倾斜校正，针对的问题正是文字图像预处理中的两个难点。

张量投票方法除了可以应用于文字特征提取和倾斜校正之外，还可应用于彩色文本的分割[11]。该方法通过张量投票使目标区域的显著性得到了增强并抑制噪声，然后通过成份标注算法，将不同的显著值标记分层，将相邻且显著值相近(小于显著值最大值和最小值之差的十分之一)的层合并后重新标注，最终使目标区域和噪声区域产生两极分化，从而达到确定分割阈值的目的。这种方法对于彩色图像用 HIS 空间的三个分量之差求和再求平均来度量，大于 15 为彩色，用色度来分析；否则判断为黑白图像，用亮度分析。

参考文献 (References)

[1] Wikipedia (2016) Optical Character Recognition. http://en.wikipedia.org/wiki/Optical_character_recognition

[2] 郭军，马跃，盛立东，等. 发展中的文字识别技术[J]. 电子学报, 1995, 23(10): 184-187.

[3] 叶齐祥. 图像和视频文字检测技术研究[D]: [博士学位论文]. 北京：中国科学院计算技术研究所, 2006: 7.

[4] 程艳芬. 离线阿拉伯手写体光学文字检测方法研究[D]: [博士学位论文]. 武汉：武汉理工大学, 2009: 32-46.

[5] 王家全, 李爱中. 部件在印刷体汉字识别中的应用[J]. 微机发展, 1998, 8(1): 17-19.

[6] 苏统华. 脱机中文手写识别——从孤立汉字到真实文本[D]: [博士学位论文]. 哈尔滨: 哈尔滨工业大学, 2008.

[7] 孙羽菲. 低质量文本图像 OCR 技术的研究[D]: [博士学位论文]. 北京: 中国科学院研究生院(计算技术研究所), 2011: 13-18.

[8] 田大增. 视觉文档图像识别预处理[D]: [博士学位论文]. 保定: 河北大学, 2007: 21.

[9] 郭平欣, 张淞芝. 汉字信息处理技术[M]. 北京: 国防工业出版社, 1985: 1.

[10] Lee, M.S. and Medioni, G. (1998) A Unified Framework for Salient Curves, Regions, and Junctions Inference. *Proceedings of 3rd Asian Conference on Computer Vision*, PRC, Springer, Hong Kong, 2, 315-322.

[11] 魏琳, 陈秀宏. 基于张量投票和成份标注的彩色文本的分割[J]. 计算机工程与设计, 2009, 30(2): 478-480.