基于中心与圆周的英文字符识别方法研究

蓝章礼
（重庆交通大学计算机与信息学院　重庆 400074）

摘　要　针对目前字符识别方法对旋转角度大的字符难以识别的问题，本文提出一种基于中心与圆周的英文字符识别方法(CCR)，通过对字符中心和适当半径圆周上的点进行分析来识别英文字符。文章通过理论分析和实验，证明了该方法能够对任意旋转角度的英文字符进行正确识别，并给出了需要进一步研究的问题。

关键词　英文字符，识别，中心，圆周

Research on English Character Recognition Method Based on Center and Circumference

LAN Zhang-Li
(School of Computer and Information, Chongqiong Jiaotong University, Chongqing 400074)

Abstract　This paper addresses an English character recognition method based on center and circumference (CCR) aiming at the problem that exiting recognition methods can't recognize inclined character. CCR recognizes English character by analyzing the center and circumference with proper radius. This paper proves CCR can recognize English character correctly which is incline with any angle by theoretical analyzing and experiment, and gives some problems needed to research further.

Keywords　English character, Recognition, Center, Circumference

1　引言

英语是当今国际通用语言，因此有大量的英文文档需要进行整理、查询、统计和电子化，而实现英文文档的自动识别可以大大提高工作效率，因而英文文档自动识别的研究具有重要意义。对英文文档的识别，通常是对单词进行识别，而单词的识别大部分是基于单字符识别的，所以单词的正确切分和字符的正确识别是提高单词识别率的关键。由于单词的切分已得到较好的解决[1~3]，因此提高英文字符的识别率是进一步改善英文文档识别性能的关键，如何利用快速有效的智能方法或技术进行字符识别成为亟待解决的问题。

为提高英文字符的识别率，许多用于英文字符的识别方法得以研究和应用，如在统计学习理论的基础上发展起来的支持向量机(SVM)[4]，特征融合及相似度判据方法[5]，基于遗传算法和BP网络的识别方法[6]，用小波理论进行字符识别[7]，隐马尔柯夫模型(HMM)[8]等。

这些方法对英文字符的识别都有一定的优点和实用价值，但是都有一个共同的缺点，即只能对正立的字符或只有小角度旋转的字符进行识别，虽然可通过其它方法对倾斜的字符进行校正，但不能同时对同一页面上倾斜角度不同的字符进行校正，如坐标系中的字符，有的正立，即倾斜角度基本为0，有的呈45度角，有的呈90度角。这就很难用现有的字符识别方法对这些倾斜角度各不相同的字符进行识别。

为此，本文提出一种基于中心与圆周的英文字符识别方法（Center and Circumference Recognition，以下简称为CCR），用于解决识别任意倾斜角度的英文字符。

2　基于中心与圆周的英文字符识别方法

对于文字和字符的识别，关键是通过提取文字或字母的特征，然后通过统计与分析，进行特征对照，进而识别文字或字符。

2.1　英文字符的形状特征分析

传统的英文字符识别，都是从水平方向、竖起方向，或者同时从水平方向和竖直方向，或者从某一特定角度提取字符的形状特征。目前比较常用的字符识别特征如方向线素特征[9]、穿越特征、网格特征、外围特征、投影特征、边缘特征、四边码特征[10]等，这些特征都是基于水平方向或竖起方向进行提取，要求字符倾斜角很小，不能对任意倾斜角度的英文字符进行识别。

通过观察和分析英语字母的特征，容易发现，英文字符除在水平和竖直方向上可以提取特征外，一些字符的中心有笔画经过，而其它字符中心没有笔画经过，如字母B的中心有笔画，而字母C的中心无笔画；当以字母的中心点为圆心，按一定的半径R画一个圆，则该圆的圆周上黑点和白点的分布比例又各不相同，如C和D，虽然中心都没有笔画，但以中心为圆心，以一定长度R为半径的圆周上，黑白点的比例又有所不同。为此，完全可以利用英文字符的这一特性进行特征的提取和识别。

2.2　中心与圆周识别方法的基本原理与步骤

中心与圆周字符识别方法(CCR)的基本原理是对字符的中心及以中心为圆心、适当半径为圆的圆周上的点进行分析，通过判断有无笔画和黑白点比例进行分析对照。当前两项结果相同或相近时，再次以中心为圆心，以先前半径的1/2为半径作一个新的圆，判断新圆周上黑白点比例，再次进行分析对照，得出结果。

其基本步骤是：

1)进行图像预处理，即将字符图片转化为二值图像并处理成固定大小的图片，去除噪声，有笔画处为黑色，标识为

蓝章礼　讲师，博士研究生，主要研究方向：计算机算法设计与研究。

“1”,无笔画处为白色,标识为“0”;

2)判断字符的中心点标识,将中心有笔画和无笔画的字符分开;

3)以中心为圆心,以一个预先设定的值为半径 R 画圆(R 可以为图片大小的一定比值),判断圆周上黑点与白点的比例,根据比例与现有数据进行对照,若为典型值,则可得出结论,识别完成,若比例结果可能发生歧义,则进行第四步;

4)以中心为圆心,取先前半径的 1/2 作为新的半径画圆,判断该圆周上黑点与白点的比例,再次进行对照,得出结论。其流程图如图 1。

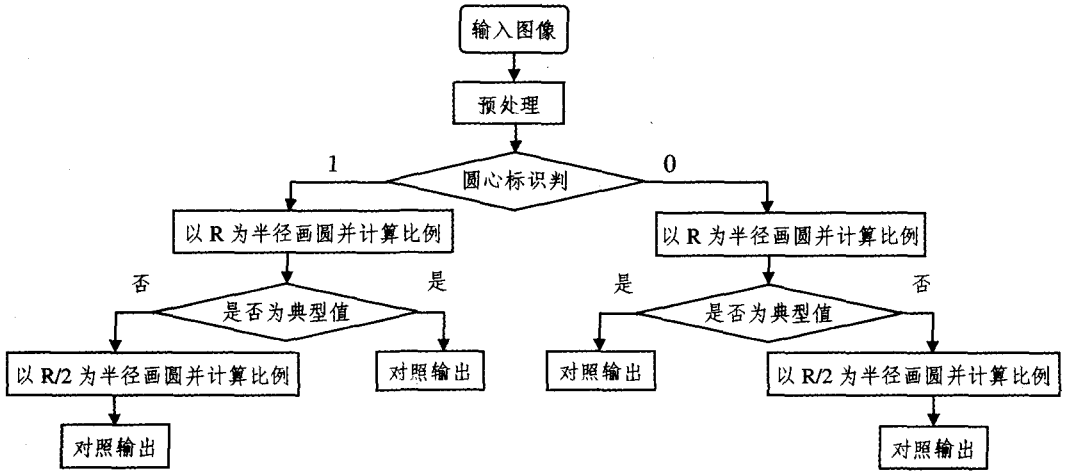

图 1　中心与圆周识别方法流程

2.3　中心与圆周识别方法的特点和优点

相比现有的英文字符识别方法,CCR 方法有以下几个特点:

1)对待识别字符的摆放角度无任何要求。因为该方法对字符进行识别时只需获取中心及以中心为圆心的圆周上的图像信息,而圆是中心对称的,与图片的倾斜角度没有任何关系,所以,在识别时对字符的摆放角度没有要求,这是区别于其它识别方法的关键特点;

2)理论上在一定范围内不需要对字符图片的大小做归一化处理。因为在识别时只对中心和圆周上黑白的比例进行运算,从理论上而言在一定范围内中心点是不变的,而画圆时可以按字符图片的大小的一定比例取半径,圆周上的黑白比例是一定的,所以理论上没有必要统一字符图片的大小,或者图片大小相近时不必做归一化处理;

3)需要判断中心是否有笔画。要准确判断中心点的位置及有无笔画经过;

4)需要运算圆周上黑点与白点的比例。

CCR 的主要优点在于:现有的字符识别方法都需要字符摆放基本正立,即倾斜角要接近于 0 度,限制了字符识别的应用范围。而 CCR 对字符摆放角度完全没有要求,这是其它识别方法所不具有的,因此,对 CCR 进行深入的研究有重要的理论意义和实用价值。

3　实验设计与结果

为验证 CCR 方法的实用效果,设计了以下实验。

将 26 个字体为 Times New Roman(未加粗)的英文字母制成大小为 96×96pixels 的二进制图片,未加入噪声,其中字符在图片中倾斜角为 0 度,字符高度为 80pixels 并居于图片正中心。对所有图片的中心进行识别,有笔画通过中心的标识为 1,无笔画通过中心的标识为 0。然后以图片中心为圆心,以 42pixels 为半径画圆,对圆周上点的黑白情况逐个进行识别,并运算黑白比例,如图 2 所示。对于字母 Q、D、E,其中心均无笔画,且圆周上黑白比例接近,不易判断,为此,再以中心为圆心,以 21pixels 为半径画圆,运算 R/2 圆周上黑白比例。同理,对中心为黑点的字母 F、N、S 以同样的方法进行处理,如图 3 所示。所有字符的处理结果如图 4 所示。图中“0”和“1”表示中心有无笔画经过,其余数字表示以 R 为圆的圆周上黑白百分比及 R/2 为圆的圆周上黑白百分比。

图 2　字母 A 的圆心及圆周黑白识别

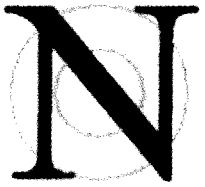

图 3　字母 N 的圆心及圆周黑白识别

完成上述工作后,按本文算法对所有字符图片进行识别,识别时参照中心及圆周上的黑白比例进行对照,对于未作旋转的字符图片识别率达到 100%。然后对图片 15 度、30 度、45 度、60 度、90 度、120 度、135 度、180 度、270 度等典型角度和其它的任意角度进行旋转后进行识别,识别时中心的黑或白要求完全匹配,圆周上的黑白比例按照最接近于图 4 中的那个比例就认为是相对应的字符。实验中发现,由于中心的寻找不够准确,在只进行一次圆周识别时容易产生误判,为此,对于一些字符进行了 R/2 圆的判断,在经过 R 及 R/2 的圆周判断后,对于实验中给定的理想图片识别率可达到 100%。另外,实验还对 Arial Black 字体的英文字符进行了识别,效果同样良好。

(下转第 249 页)

2000/NOTE-SOAP-attachments-20001211
3 Message Transmission Optimization Mechanism. http://www.w3.org/TR/2005/REC-soap12-mtom-20050125
4 The SSL Protocol. http://wp.netscape.com/eng/ssl3/ssl-toc.html
5 HTTP Secure. http://www.ietf.org/rfc/rfc2818.txt
6 XML-Signature Syntax and Processing. http://www.w3.org/TR/xmldsig-core/
7 XML Encryption Syntax and Processing. http://www.w3.org/TR/xmlenc-core/
8 XML Key Management Specification (XKMS). http://www.w3.org/TR/xkms/
9 Web Services Security. http://www-128.ibm.com/developerworks/webservices/library/ws-secure
10 Web Service Security SOAP Messages with Attachments Profile. http://www.oasis-open.org/committees/download.php/10902/wss-swa-profile-1.0-&e=7152
11 Bosworth A, Box D, Gudgin M, et al. Xml, SOAP and Binary Data. http://www.xml.com/pub/a/2003/02/26/binaryxml.html
12 Web Services Reliable Messaging Protocol (WS-ReliableMessaging). http://xml.coverpages.org/WS-ReliableMessaging200502.pdf
13 黄涛,陈宁江,魏俊,等. OnceAs/Q:一个面向 QoS 的 Web 应用服务器. 软件学报,2004,15(12):1787～1799
14 the Axis Development Team. Axis Architecture Guide. http://ws.apache.org/axis/
15 Ford W, Baum M S. Secure Electronic Commerce. Prentice Hall PTR, 1997

(上接第 242 页)

实验证明,利用本文的 CCR 方法在理想情况下完全可以正确标识 26 个英文字符,而不管这些字符是否存在倾斜,达到了对倾斜英文字符的识别效果。

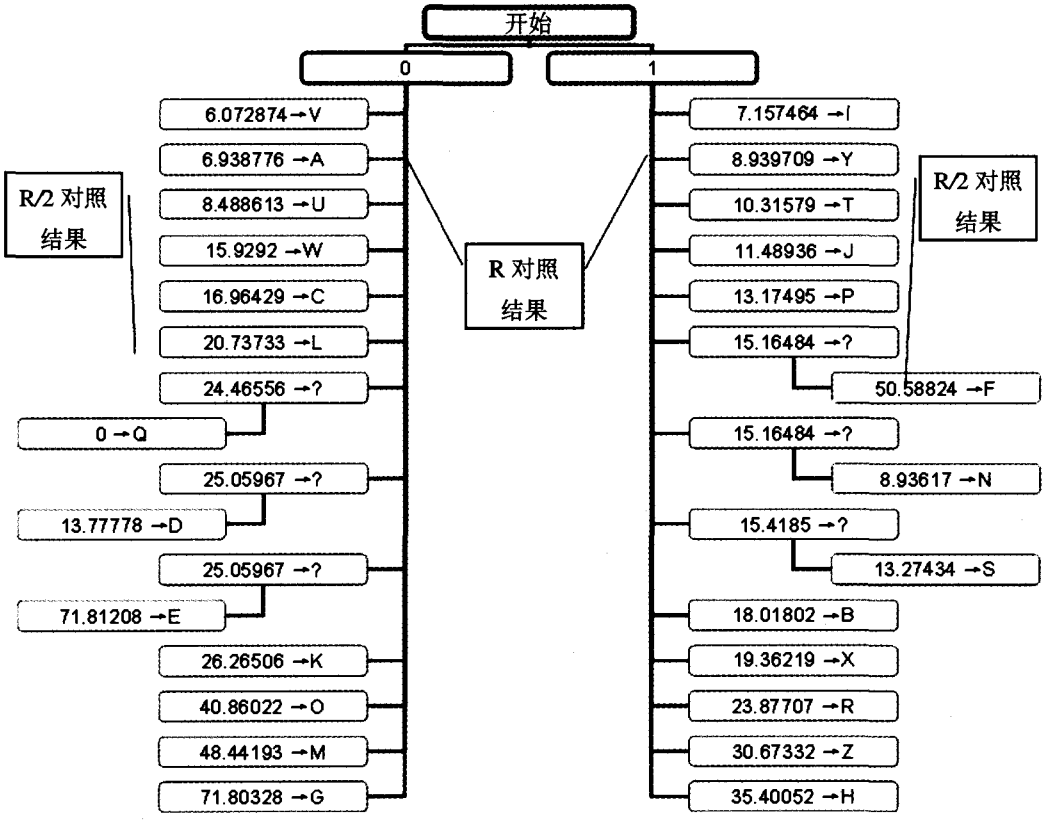

图 4　26 个英文字母的圆心及圆周黑白比例

结论与进一步的工作　通过理论分析与实验,可以认为基于中心与圆周的英文字符识别方法是一种有效的字符识别方法,理论上讲它不仅适用于各种字体的英文字符识别,对于其它字符的识别也应该同样有效,有必要进行更加深入的理论研究和实用化。

但是,还有一些重要的问题需要进一步研究和更好的解决,首先是中心的判定问题,虽然现在有一些中心(圆心)快速寻找算法,但精度和速度还不够高,特别是判断精度还有待提升,而中心是否找得准确,将直接影响识别的结果,因此有必要研究更好的中心寻找算法;二是半径问题的研究,即如何确定比较合适的半径值 R,什么时候需要进行 R/2 判定,是否会出现 R 与 R/2 都无法识别的情况,是否还需要 3R/4 判定等;三是图像的预处理,如果图片质量不够好,识别率将会下降,怎样提高图片的质量,图片质量要达到什么程度效果才好;四是研究如何将基于中心与圆周的识别方法用于数字、汉字乃至其它文字的识别;五是研究如何将该方法与其它方法相结合,提高识别的准确率。

参 考 文 献

1 Vineiarelli A. A Survey on Off-line Cursive Word Recognition [J]. Pattern Recognition, 2002, 35: 1433～1446
2 Bozinovie R M, Srihari S N. Off-Line Cursive Script Word Recognition [J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 1989, 11(1): 68～83
3 武振军,丁晓青. 鲁棒的多体印刷英文识别系统的实现[J]. 计算机工程与应用,2001,37(20):120～122
4 Vapnik V. Statistical Learning Theory[M]. New York: Wiley, 1998
5 吴锐,赵巍,尹芳,唐降龙. 特征融合及相似度判据在英文识别中的应用[J]. 计算机工程与应用,2005,16:55～57
6 苗琦龙,栾新. 基于遗传算法和 BP 网络的文字识别方法[J]. 计算机应用,2005,25(12):330～332
7 沈会良,李志能. 基于矩和小波变换的数字、字母字符识别研究[J]. 中国图象图形学报,2005,5(3):249～252
8 Hassin A H, et al. Printed Arabic Character Recognition Using HMM [J]. J Comput Sci Technol, 2004, 19(4): 538～54
9 张忻中. 汉字识别技术[M]. 清华大学出版社,1992
10 Kato N, et al. A Handwritten Character Recognition System Using Directional Element Feature and Asymmetric Mahalanobis Distance[J]. IEEE Tran on PAMI, 1999, 21(3): 258～262