小写金额的多模式切分与识别算法

谷军霞 丁晓青
(清华大学电子工程系,北京 100084)

摘 要 针对带表格的中文支票小写金额的自动识别问题,提出了一种多模式切分和识别算法。根据小写金额不同部分的切分和识别难度,采取了3种递进的模式:预切分模式、连写0检测模式和基于识别的切分模式。其中预切分模式用来处理小写金额中不粘连的单字;连写0检测模式用来检测并识别连写的0;基于识别的切分模式用来处理非连写0的粘连部分,在这个模式中采用了遗传算法来加速最优解的搜索过程。利用从银行采集的1 053张真实支票样本进行测试,在拒识率为33.6%时,小写金额串的识别率达到66.1%,实验结果证明这种算法可以提高真实支票小写金额的识别率。

关键词 多模式切分 识别 小写金额 支票
中图法分类号:TP391.43 **文献标识码**:A **文章编号**:1006-8961(2008)04-0696-06

Multi-model Segmentation and Recognition Algorithm of Courtesy Amount

GU Jun-xia, DING Xiao-qing
(*Department of Electronic Engineering, Tsinghua University, Beijing* 100084)

Abstract A multi-model segmentation and recognition algorithm of courtesy amount on Chinese bank checks with the form lines is presented in this paper. Based on some characteristics of Chinese bank checks, we adopt three models for different parts of the courtesy amount. The pre-segmentation model deals with the isolated characters. The touching zeros detection model is designed for the part of touching zeros. The segmentation-based recognition model deals with other touching part in the courtesy amount. In the third model, we use genetic algorithm to accelerate the searching process. The system is validated with 1 053 real bank checks. The reject rate is 33.6%, and the recognition rate at the amount level can reach 66.1%. The experiment results show that the recognition rate of the real bank checks can be improved with this new method.

Keywords multi-model segmentation, recognition, courtesy amount, bank check

1 引 言

支票小写金额的自动录入已经成为模式识别一个很活跃的应用领域[1]。但是支票的复杂和多变使得这项工作极富挑战性。由于每个人书写风格不同,并且手写字符经常覆盖、粘连,还存在噪声污染、断笔的情况,同时数字本身的笔画很少,从而使得在切分过程中容易造成过切分和欠切分的情况,这些都将直接影响单字识别精度。而只要有1个数字识别错误,整张支票的识别就宣告失败。所以小写金额准确切分与识别的问题一直以来都是支票识别研究的重点[1,2]。

按照切分与识别是否有关,小写金额的切分识别算法可以分为两类:切分识别无关算法(segmentation-then-recognition, STR)[2]与切分识别结合算法(segmentation-based-recognition, SBR)[1]。这两种策略各有优缺点,识别信息对于切分具有很重要的作用,它可以纠正切分的错误,而STR算法没有利用这一点,当真实的切分路径与预先假设不相符时就会出现切分错误。SBR算法充分利用识别的信息,但是切分的性能要受到单字分类器性能的影响,当

收稿日期:2006-10-13;**改回日期**:2006-12-12

第一作者简介:谷军霞(1981~),女。现为清华大学电子工程系直博研究生。研究领域包括文字切分与识别、视频分析。
E-mail:gujx04@mails.tsinghua.edu.cn

识别核心得到的识别分数与实际结果不相符时也会影响切分的性能。

中文支票上的小写金额具有很多特性,充分利用这些性质可以帮助我们提高识别的性能。中文支票上一般都有表格线,表格线一方面会干扰要识别的字符串,另一方面也会对切分提供非常有用的信息。中文小写金额还有一些简单的语法规则,例如人民币符号“￥”只可能出现在首位。还有一个重要特性就是不同支票甚至同一张支票的不同部分往往具有不同的切分复杂度。如图 1 所示,没有被表格线干扰、没有粘连的字符就很容易被切分识别,但是被表格线干扰或者互相粘连、覆盖、越格的字符串部分就很难正确切分和识别,因此对具有不同切分难度的部分要采用不同的切分模式。

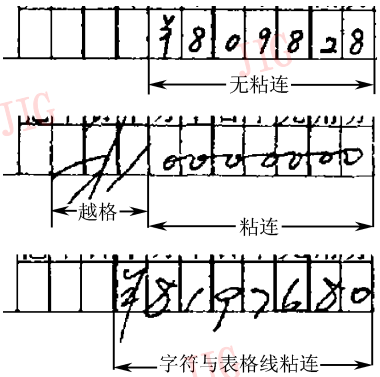


图 1　复杂多变的小写金额

Fig. 1　Variable courtesy amounts

基于中文支票小写金额的这些特性,提出了多模式切分和识别算法(multi-model segmentation and recognition algorithm, MSRA)。这个算法融合了 STR 和 SBR 这两种识别策略的优点,即不根据识别信息就可以确保切分正确时,采用 STR 策略;在不能保证切分正确时,采用 SBR 策略搜索最优切分路径,从而进一步提高了小写金额切分识别的精度。

2　中文支票识别系统概述

在现有实际应用的银行支票识别系统中,往往采用大小写金额融合识别的算法[3~6]。整个中文支票识别系统也采用了这类算法,其框图如图 2 所示。

本文重点在于小写金额识别算法。整个系统的其他部分可以参见文献[6]。

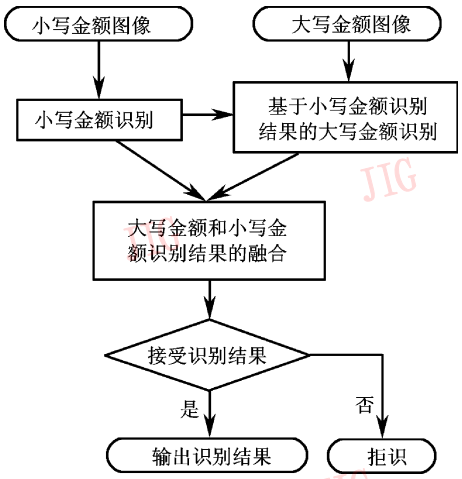


图 2　中文支票识别系统

Fig. 2　Recognition system of Chinese bank checks

3　小写金额的多模式切分识别算法

中文支票小写金额的典型样本如图 3 所示。它由 3 部分组成:表格线、￥符号、数字串。将￥符号和数字串作为要识别的字符串。

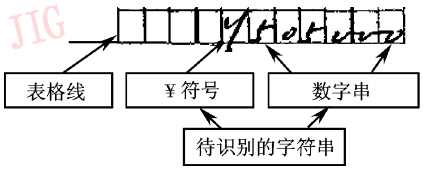


图 3　小写金额的典型样本

Fig. 3　Typical sample of courtesy amount

MSRA 识别系统的框图如图 4。此系统的输入是二值化的小写金额图像,输出是小写金额字符串的识别结果。

3.1　预处理模块

预处理模块用来完成表格线的检测和去除。表格线检测主要利用了投影法,并结合先验信息进行进一步精细地修正;表格线的去除采用了笔画保护措施。

3.1.1　表格线的检测

因为支票小写金额部分的表格线都是水平、垂直线或者近似的水平、垂直线,所以采用了投影法检测表格线。但是并不是所有的样本都能通过投影检测到精确的表格线位置,有一些样本由于倾斜使得水平线检测不准确,还有一些样本由于字符串竖直笔划的影响使得垂直线检测不精确,因此需要进行表格线的修正。

假设输入支票的表格大小是固定的,令 w_1 表示

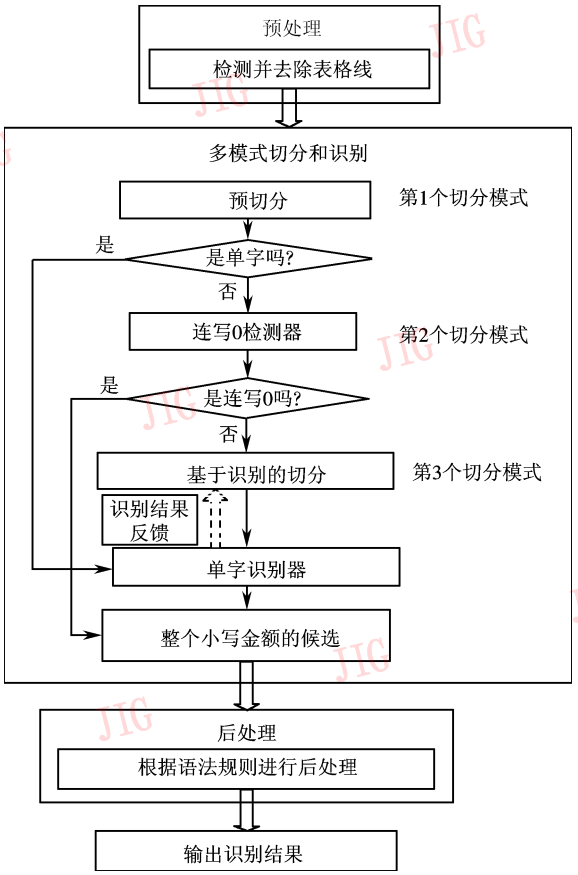


图 4 MSRA 算法的系统框图

Fig. 4 The chart of the MSRA

表格的宽度，w_L 表示表格线的宽度，w_T 和 w_L 可由训练集中支票样本的表格宽度和表格线宽度的平均值获得。当用投影法检测到的水平表格线宽 W 大于 w_L 时，表明样本倾斜了。计算倾斜角的示意图如图 5 所示，其中，L 表示水平表格线的长度，则倾斜角的大小 α 可由下式得到：

$$\alpha = \arctan\left(\frac{W - w_L}{L}\right) \tag{1}$$

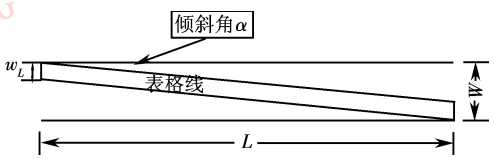


图 5 倾斜角计算示意图

Fig. 5 Sketch map of skew angle

字符竖直笔划对于垂直表格线的检测来说是噪声，会影响垂直线检测的精度。可以根据 w_T 进行纠正。一般而言，左起第 1 条垂直表格线一般不会受到垂直笔划的影响，因此可以作为一个基准。其他表格线的位置可以由下式大致估计出来：

$$L_i = L_{i-1} + w_T \quad i > 1 \tag{2}$$

式中，L_i 表示第 i 条垂直表格线的估计位置。

3.1.2 表格线的去除

表格线的去除是根据表格线宽采取笔画保护策略而进行的。首先采用文献[7]的方法估计待识别字符的平均笔画宽度 λ。然后在表格线周围标记与表格线垂直方向的游程[8]，去掉与表格线宽近似的游程，对于大于表格线宽的游程，根据估计得到的笔画宽度去除部分表格线。

去除表格线后，用形态学运算消除小的噪声，并闭合断裂的笔画[9]。

3.2 多模式切分和识别模块

预处理后的一个典型样本如图 6 所示。根据切分和识别的难度不同，这个字符串可以分为 3 部分：孤立的单字、粘连的 0、其他粘连部分。采用相应的 3 种模式来处理这些不同的部分。

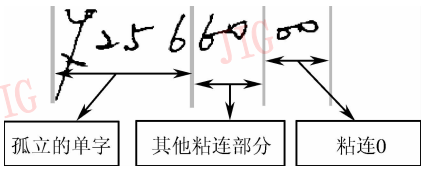


图 6 小写金额待识别字符串的典型样本

Fig. 6 Typical sample of character string of courtesy amount

3.2.1 单字符分类器

在整个系统中使用了 *MQDF*（*modified quadratic discriminant functions classifier*）单字分类器，并提取了多方向线素特征[10]。该分类器输出的是字符的识别距离，利用一个模糊函数将识别距离映射为字符的后验概率[9]，如下式：

$$p(\boldsymbol{c}/\boldsymbol{s}) = e^{-d(\boldsymbol{s},\boldsymbol{c})}, \boldsymbol{c} \in \boldsymbol{C} \tag{3}$$

式中，$\boldsymbol{C}$ 表示字符集，$\boldsymbol{s}$ 表示单字符图像，$\boldsymbol{c} \in \boldsymbol{C}$ 是 $\boldsymbol{s}$ 的识别候选，$d(\boldsymbol{s},\boldsymbol{c})$ 为分类器输出的识别距离，$p(\boldsymbol{c}/\boldsymbol{s})$ 表示后验概率。

3.2.2 预切分模式（STR 与 SBR 策略相结合）

预切分模式主要用来切分并识别小写金额中孤立的单字。

首先对字符串进行欠切分，设 d_{ij} 表示字符串中相邻两个连通分量 i,j 的最近距离，α 为一正常数，λ 为平均笔画宽度。若 d_{ij} 满足下式：

$$d_{ij} > \alpha\lambda \tag{4}$$

并且连通分量 i,j 之间存在一条表格线，则在这两个

连通分量之间进行切分,如图 7 所示。

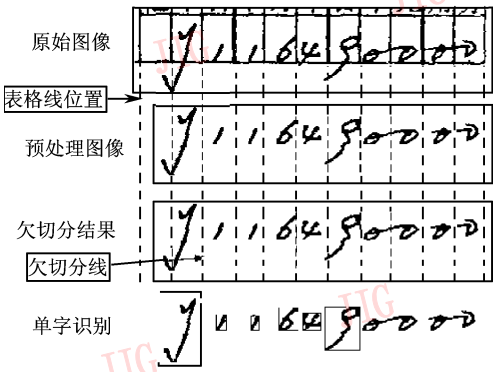


图 7　预切分模式示例

Fig. 7　Sketch map of the pre-segmentation

然后判断输出的欠切分子图像是否为单字。其规则为占用一个表格的子图像是单字(STR 策略),如果左边的第 1 个子图像占有多个表格(可能是“¥”符号占用了多个表格),就送入单字识别器,根据识别器输出的识别距离判断该子图像是否是“¥”符号(SBR 策略)。

最后保存判断为单字的子图像的识别候选,非单字子图像送入连写 0 检测模式。

3.2.3　连写 0 检测模式

预切分之后非单字子图像中包含大量的连写 0,在 2 900 个训练支票样本中,连写 0 子图像约占所有非单字子图像的 60% 以上。并且连写 0 要是被切分为单个数字就很容易识别错误。因此有必要对连写 0 进行整体的检测和识别。

小写金额中连写 0 的个数从 2 到 10 不等,不容易采集如此多种的样本,因此连写 0 检测器的核心是双 0 检测器,即对于占据多于两个表格的非单字子图像按照表格线预先切分为双字子图像,如图 8 所示。其中连写字符串中间的数字会在两个子图像中出现,只有这两个子图像都被识别为双 0,才认为这个数字是 0。

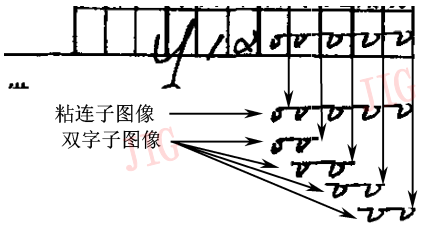


图 8　双字子图像的获取

Fig. 8　Double characters

双 0 检测器提取的特征与单字分类器一样,都是方向像素特征,因为双 0 检测器是个二类分类器,只需要判断输入的双字子图像是不是双 0,所以采用了泛化效果比较好的 Fisher 分类准则。

3.2.4　基于识别的切分模式:表格分配算法(SBR 策略)

这个切分模式用来处理非连写 0 的粘连部分和互相覆盖部分,这部分子图像在整个小写金额中所占的比例已经很少了,不到 10%。在一般的基于识别的粘连数字串切分算法中,首先进行过切分,然后根据识别的结果选择合适的切分路径[9]。而待切分识别的字符串有了表格的限制,即要识别的字符串长度已经知道了,并且大致分布的位置也知道了,利用这两个先验知识来提高切分正确率,因此采用了基于识别的表格分配算法[11],即将过切分后的连通分量分配到各个表格中,之后将每个表格内的连通分量作为一个字符进行识别。假设字符串占据的表格数为 N,该算法的步骤如下:

(1) 根据文献[9]的方法对输入字符串进行过切分,过切分的候选点除了投影分析和轮廓分析得到的点之外还包括字符笔画与垂直表格线的交点。过切分后的连通分量个数设为 $M(M \geqslant N)$。

(2) 为每个表格分配一个主连通分量(main connected component,MCC)[11]。分配的原则是取该表格内面积最大的连通分量作为该表格的 MCC。其他没有被选为 MCC 的连通分量称为 non-MCC,共有 $(M-N)$ 个 non-MCC。

(3) 为 non-MCC 选择初始表格分配。将距该 non-MCC 最近的 MCC 所属的表格定为该 non-MCC 的初始表格,记为 $\{T_i^0\}$,$1 \leqslant i \leqslant (M-N)$。

(4) 根据识别的结果得到 non-MCC 的最佳表格分配。假设第 i 个 non-MCC 有 4 种分配情况:第 1 种是属于初始表格 T_i^0;第 2 种是属于初始表格左边的表格 (T_i^0-1);第 3 种是属于初始表格右边的表格 (T_i^0+1);第 4 种是噪声。那么对于有 $(M-N)$ 个 non-MCC 的字符串来说,最多的分配情况有 4^{M-N} 种,随着 $(M-N)$ 呈指数增长,所以采用了遗传算法(genetic algorithm,GA)来加速搜索的过程。

(5) 将每一个表格内的连通分量作为一个字符,由单字分类器进行识别,输出识别结果的候选。

其中,将 GA 算法用于搜索 non-MCC 最优分配的关键是适应度函数的构造。适应度函数由识别适应度和几何适应度两部分组成,其定义如下式:

$$f(\boldsymbol{D}) = \beta \boldsymbol{r} + (1-\beta)\boldsymbol{g} \tag{5}$$

式中，$\boldsymbol{D}$ 表示一种 non-MCC 的解空间，$\boldsymbol{r}$ 表示归一化的识别适应度，$\boldsymbol{g}$ 表示归一化的几何适应度，$\beta(0 \leqslant \beta \leqslant 1)$ 是加权因子。

归一化的识别适应度 r 的定义如下：

$$r = \frac{1}{N}\sum_{i=1}^{N} p(c_i/s_i) \tag{6}$$

式中，N 表示字符串的长度，也是这串字符所占用的表格个数；c_i 表示第 i 个表格字符识别结果的第 1 候选；s_i 表示第 i 个表格的图像。

归一化的几何适应度 g 的定义如下：

$$g = \frac{1}{N}\sum_{i=1}^{N} p(w_i/c_i) \tag{7}$$

式中，w_i 表示字符 c_i 的宽；$p(w_i/c_i)$ 表示字符 c_i 宽度的概率分布，可由训练集的直方图得到。

用 $\{-1,0,1,2\}$ 表示 non-MCC 的 4 个状态，则解空间 $\boldsymbol{D}$ 可以表示为 $\{(d_1,\cdots,d_{M-N})\}$，其中 $d_i \in \{-1,0,1,2\}$，$1 \leqslant i \leqslant (M-N)$。

这个算法使用于 non-MCC 比较多的情况，对于 non-MCC 比较少的情况，可以用全搜索算法代替 GA 算法。

3.3 后处理模块

单字识别器为每个单字提供了 5 个候选，可以利用这些候选值，根据语法规则对识别结果进行后处理。这些语法规则主要有：

（1）￥符号只可能出现在左边的第 1 个字符，如果出现在其他位置则用下一个候选代替；

（2）最高位的数字不能是 0，如果是 0，则用下一个候选代替。

在整个中文支票的识别系统中，小写金额识别的结果还会与大写金额切分识别融合以得到整张支票金额的识别结果。

4 实验结果

训练集和测试集都来自实际的银行支票，其中训练集包括 2 900 个支票样本，有单字符 20 724 个，用来训练算法中的参数、单字分类器及双 0 分类器。测试集包括 1 053 个支票样本。

采用了拒识率、识别率和接受样本的识别错误率[3] 3 个准则来评价本文算法的性能。这 3 个评价准则的定义如下：

$$\begin{cases} R_{\text{reject}} = \dfrac{N_{\text{reject}}}{N_{\text{total}}} \times 100\% \\ R_{\text{recognition}} = \dfrac{N_{\text{right}}}{N_{\text{total}}} \times 100\% \\ R_{\text{substitution}} = \dfrac{N_{\text{error}}}{N_{\text{total}} - N_{\text{reject}}} \times 100\% \end{cases} \tag{8}$$

式中，N_{total}、N_{reject}、N_{error}、N_{right} 分别表示支票样本的总个数、拒识样本的个数、识别错误的样本个数、识别正确的样本个数。

与大写金额融合之后，整个支票识别系统的性能及与其他算法的比较如表 1 所示。由于没有统一的支票识别测试样本，表 1 所示的比较结果只是一个很粗略地相对实验结果，有的系统测试样本数目过小，测试结果的准确度也受影响。

图 10 是 $R_{\text{recognition}}$ 随 $R_{\text{substitution}}$ 的变化曲线。根据文献[3]，当 $R_{\text{substitution}}$ 在 0.1% 左右时，6 个不同国家（Australia，Ireland，Canada，England，USA，France）的支票识别率都低于 50%，而在本文算法当 $R_{\text{substitution}}$ 在 0.1% 左右时，识别率在 60% 左右。

图 11 是一些典型的错误样本。小写金额切分的错误主要是由于人民币符号“￥”被过切分造成的；单字识别错误主要是相似字造成的。

表 1 识别性能比较

Tab. 1 Comparison of the performance

算法	R_{reject}(%)	$R_{\text{recognition}}$(%)	$R_{\text{substitution}}$(%)	测试样本总数	测试样本来源
文献[1]算法：小写金额独立识别	31.0	68.5	0.5	400	加拿大支票
文献[2]算法：小写金额独立识别	0.0	89.3	10.7	84	中文支票
文献[4]算法：大小写金额融合识别	56.2	43.8	0.0	235	法国支票
文献[5]算法：大小写金额融合识别	3.30	95.96			韩国支票
文献[3]算法：大小写金额融合识别	42.8	57	0.4	10 957	法国支票
本文算法：小写金额独立识别	0.0	86.5	13.5	1 053	中文支票
本文算法：大小写金额融合识别	33.6	66.1	0.4	1 053	中文支票

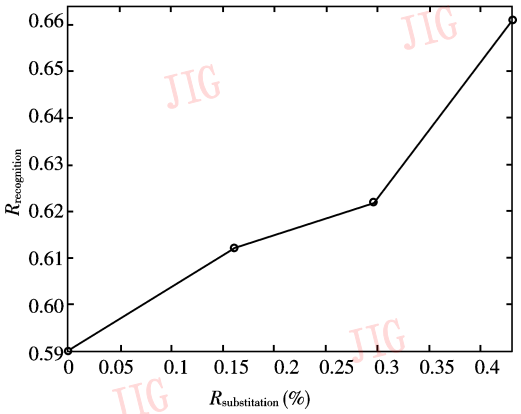
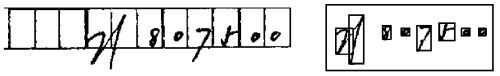
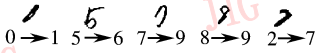


图 10　识别率变化曲线

Fig. 10　The curve of the recognition rate



(a)切分错误(左边是原始图像,右边是“¥”过切分的结果)



(b)识别错误的单字

图 11　错误样本示例

Fig. 11　Error samples

5　结　论

提出了一种支票小写金额的多模式切分和识别算法。该算法基于中文支票小写金额的特点,将小写金额根据切分难度分为 3 部分,用递进的 3 种切分策略来处理。实验结果表明这种算法可以使得小写金额的识别率达到比较高的水平。

参考文献(References)

1　Zhang L Q, Suen C Y. Recognition of courtesy amounts on bank checks based on a segmentation approach[A]. In: Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition[C], Ontario, Canada, 2002: 298 ~ 302.

2　Song Hong-dong, Wang Ya-li, Xia Shao-wei. Recognition system of handwritten numerals on checks using support vector machine[J]. Computer Engineering and Applications, 2003, (3): 58 ~ 60. [宋鸿冬,王亚利,夏绍玮. 基于支撑向量机的支票手写体数字识别系统[J]. 计算机工程与应用. 2003, (3): 58 ~ 60.]

3　Nikolai Gorski, Valery Anisimov, Emmanuel Augustin, *et al*. Sergey maximov industrial bank check processing: the A2iA CheckReader™ [J]. International Journal on Document Analysis and Recognition, 2001, **3**(4): 196 ~ 206.

4　Dzuba G, Filatov A, Gershuny D, *et al*. Check amount recognition based on the cross validation of courtesy and legal amount fields [J]. International Journal on Pattern Recognition and Artificial Intelligence, 1997, **11**(4): 639 ~ 655.

5　Tae-Chang Jee, Eun-Jin Kim, Yillbyung Lee. Error correction of Korean courtesy amount in back slips using rule information and cross-referencing[A]. In: Proceedings of the International Conference on Document Analysis and Recognition [C], Bangalore, India, 1999: 95 ~ 98.

6　Gu Jun-xia, Ding Xiao-qing. Fusion recognition of courtesy and legal amounts on Chinese handwritten bank checks [A]. In: Proceedings of the International Conference on Signal Processing[C], Guilin, China, 2006: 1738 ~ 1741.

7　Garris M D. Component-Based Handprint Segmentation Using Adaptive Writing Style Model [R]. NIST Internal Report 5843, Gaithersburg, USA: National Institute of Standards and Technology, 1996.

8　Yi-Hong Tseng, Hsi-Jian Lee. Interfered-character recognition by removing interfering-lines and adjusting feature weights [A]. In: Proceedings of the fourteenth International Conference on Pattern Recognition [C], Brisbane, Austrialia, 1998: 1865 ~ 1867.

9　Lei Yun, Liu C S, Ding X Q, *et al*. A recognition based system for segmentation of touching handwritten numeral strings [A]. In: Proceedings of the ninth International Workshop on Frontiers in Handwriting Recognition [C], Tokyo, Japan, 2004: 294 ~ 299.

10　Liu Hai-long, Ding Xiao-qing. Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes [A]. In: Proceedings of the International Conference on Document Analysis and Recognition[C], Seoul, Korea, 2005: 19 ~ 23.

11　Qian Wei-wei, The Back Check Recognition System [D]. Beijing: Tsinghua University, 2002. [钱伟威. 支票识别系统[D]. 北京:清华大学, 2000.]